⑫ # EUROPEAN PATENT APPLICATION

㉛ Applicant: **CETUS CORPORATION**
1400 Fifty-Third Street
Emeryville California 94608(US)

㉒ Inventor: **Shoemaker, Sharon Payne**
3173 Burbank Drive
Fairfield California 94533(US)

㉒ Inventor: **Gelfand, David Harrow**
6208 Chelton Drive
Oakland California 94608(US)

㉒ Inventor: **Innis, Michael Alan**
3133 Carlsen Street
Oakland California 94602(US)

㉒ Inventor: **Kwok, Shirley Yee**
611 Lomond Circle
San Ramon California 94583(US)

㉒ Inventor: **Ladner, Martha Baillie**
2800 Barrett Avenue
Richmond California 94804(US)

㉒ Inventor: **Schweickart, Vicky**
707 San Carlos Avenue
Albany California 94706(US)

㉔ Representative: **Vossius Vossius Tauchner Heunemann Rauh**
Siebertstrasse 4 P.O. Box 86 07 67
D-8000 München 86(DE)

�54 Recombinant fungal cellobiohydrolases.

㊼ Methods and materials for obtaining recombinant forms of fungal cellulase enzymes are disclosed. An uninterrupted coding sequence for a desired cellulase can be obtained by manipulation of the genomic DNA of a cellulolytic fungus, and ligated into appropriate vectors for expression as mature, glycosylated, functional active proteins in yeast.

EP 0 137 280 A1

RECOMBINANT FUNGAL CELLOBIOHYDROLASES

Technical Field

The present invention relates to the field of cellulose degradation and modification and, more specifically, to the particular
5   enzymes useful therein. Specifically, the invention relates to the cloning and expression of the genes encoding the various enzymes involved in cellulolysis, which include endoglucanases, cellobiohydrolases and β-glucosidases derived from cellulolytic fungi.

Background Art

10   The world's most abundant renewable resource is cellulose, a polymer of glucose, which is a major structural component of wood and herbaceous materials. As a result, cellulose is a significant part of so-called waste products, such as municipal solid waste, used paper products, and by-products of agricultural crops. It would be
15   advantageous to harvest the glucose from these sources in an efficient manner, either for use directly as food, or as a starting material for other useful materials such as ethanol. Limited techniques are available for converting the cellulose starting material to a glucose end product. Various non-enzymic processes such as acid hydrolysis,
20   have been attempted, but these lack sufficient specificity and purity of product to be economically competitive.

An important alternative is the use of cellulolytic enzymes derived from natural sources, or direct use of the microorganisms which harbor these enzymes. Because the level of enzymic activity
25   available is controlled by factors which are not readily manipulable, these methods have also suffered from economic drawbacks.

Accordingly, there is an interest in developing improved strains of cellulase-producing microorganisms, and in introducing cellulase genes in recombinant form into microbial organisms which
30   otherwise are incapable of utilizing cellulose as a carbon source. In particular, it would be desirable economically to introduce cellulolytic enzyme genes into fermentation organisms, such as

fermentation yeasts, which may not naturally have the capacity to hydrolyze cellulose to utilizable substrates such as glucose, but which may have superior growth and production properties.

The typical cellulase system contains three major types of enzymes which act cooperatively in degrading cellulose (Gritzal, M., et al., Hydrolysis of Cellulose: Mech. of Enzymatic and Acid Catalyses, Brown, R. D., et al., ed. (1979) Am. Chem. Soc., Washington, D.C., p. 237; Mandels M., et al., Process Biochem. (May, 1978), p. 7). These enzyme types are cellobiohydrolases, endoglucanases, and β-glucosidases.

Cellobiohydrolases (CBH's), attack cellulose from the non-reducing ends of the cellulose polymers and yield cellobiose (a glucose dimer) as a major product. Two types of CBH enzymes are known: cellobiohydrolase-I (CBHI) which generates cellobiose exclusively, and cellobiohydrolase-II (CBHII) which generates a mixture of cellobiose and glucose. CBH enzymes have not been found in cellulolytic, non-filamentous bacteria, and thus appear to be limited to fungal and filamentous bacterial sources. They are highly specific as to substrate, and have high substrate affinity. (Derivatized forms of cellulose are not cleaved extensively by CBH).

Endoglucanases (EG's), hydrolyze internal glycosidic linkages within the cellulose chain. They are found in bacteria as well as fungi, and have a broader substrate specificity than CBH enzymes. They attack β-glycosidic linkages in a semi-random manner, and will utilize as substrates derivatized cellulose (such as carboxymethyl cellulose) and will hydrolyze 1,3,β,D- and 1,6-β,D- as well as 1,4,β,D- linkages. There are several known forms of EG enzymes which have been derived from bacteria and fungi. Endoglucanase-I (EGI) is produced in relatively high levels (5-10% of extracellular protein) in the fungus Trichoderma reesei; a second EG activity in this fungus, endoglucanase-II (EGII), has been shown to attack linear 1,4-β,D-mannan to generate primarily mannobiose.

β-glucosidases, like the CBH's, attack the non-reducing end of glycosides, but act only on solubilized cello-oligosaccharides

(chain length less than 6), to generate glucose units. The β-glucosidases thus degrade the cellobiose product of CBH's and EG's to glucose. The presence of β-glucosidases in a total cellulase system is significant not only because it provides the final end product, but because some CBH's and EG's are inhibited by cellobiose. Providing an

5 additional enzyme to deplete the concentration of this product thus aids conversion of cellulose to glucose.

The above distinct reaction specificities of the cellulase system components suggest that the enzymes can be used individually or in selected combinations to achieve controlled and limited hydrolysis

10 of cellulosic materials. The purpose of limited hydrolysis would be to alter the structural and/or functional characteristics of cellulosic materials in desired ways. Exemplary structural changes include those affecting chain length, viscosity and cellulose binding. Exemplary functional charges include those affecting

15 texture, flavor, and water binding to the material. In addition, the endoglucanases, because of their broader substrate specificity, act on other β-linked polymers. One example of this is the action of EGI on xylan. Another example is the action of EGII on linear 1,4-β,D-mannan. In experiments conducted in support of the present invention,

20 it was shown that EGII converts β-mannan to generate primarily mannobiose. This may be significant to the coffee industry, as β-mannan is a major polysaccharide coating the outside of coffee beans, and inhibits the extraction of flavor components and imparts a bitter flavor in the manufacture of coffee, particularly freeze-dried coffee.

25 The enzymes of the cellulase system have been widely studied and, indeed, attempts have been made to clone and express the coding sequences. See, for example, Whittle, D. J., et al., Gene (1982) 17:139; Cornet, P., et al., FEBS Micro. Biol. Letters (1983) 16:137; Tuse, D., Genetic Technol. News (1981) November, p. 6; Abstract No.

30 H19 ASM Annual Meeting, March 1981; Symposium ASM Annual Meeting, March 1983; Wilson, D. B., et al., Bio/Technology (1983) 1:594; Teeri, T., et al., Bio/Technology (1983) 1:696; and Collmer, A., et al., Bio/Technology (1983) 1:594-601. None of the foregoing attempts, although using various bacterial and fungal sources, including T.reesei; as

the source for the coding sequences, have reported the expression of secreted, glycosylated and biologically active fungal cellulases in yeast. Indeed, the only cellulase genes reported above to be expressed in a heterologous host are those encoding bacterial, not
5   fungal, endoglucanase.

Disclosure of the Invention

     This invention provides recombinant forms of cellulase enzymes which are related to those derived from fungal sources. The coding sequences from the various cellulase enzymes of fungal sources
10 are retrievable in uninterrupted form by the general procedures disclosed herein. By introducing such sequences in suitable expression vectors, and introducing the recombinant expression vectors obtained into suitable host organisms, large or increased amounts of these enzymes can be produced either
15 individually or in combination. Production of large amounts of such enzymes permits industrial application of their activity to hydrolyze cellulose by carrying out the reaction in vitro using these enzymes either in solubilized or in immobilized form. The enzymes thus produced are advanta-
20 geously secreted into the medium.

     According to one aspect of the invention, cellulase genes are isolated from a suitable fungal source, such as T. reesei, and their sequences determined. Using these sequences and the information they provide, it is possible to construct intron-free coding sequences ·
25 which encode either the mature enzymes alone, or, if desired, the mature enzyme in reading frame with the leader sequences which are responsible for secretion from the native host. The coding sequence is placed in a novel expression vector which allows gene expression of the heterologous sequence in foreign hosts such as
30 yeast. The host which is transformed with the expression vector is capable of producing and secreting the cellulase in mature, glycosylated, functionally active form.

According to another feature of the invention, the recombinant vectors used in transforming host organisms contain cellulase signal sequences which are shown to function in yeast for the efficient processing and secretion of cellulases from yeast. These sequences could also be used for the secretion of other proteins from yeast, preferably for the secretion of proteins that are normally secreted by their native host. Examples of such proteins include amylases, proteases, interferons, lymphokines, insulin, and hormones.

In still another aspect of the invention, an organism which produces a number of different cellulase enzymes in defined ratios may be manipulated to inactivate or modify one or more of the cellulase enzymes either to enhance the enzyme in a desired fashion or to prevent production of the enzyme. For example, in a fungal organism which produces both CBHI and CBHII enzymes, the CBHII gene may be inactivated to yield an organism which produces only the CBHI cellobiohydrolase. The method for inactivating a selected cellulase gene generally includes inactivating (e.g., by a large-segment deletion) the recombinant cellulase gene constructed according to the invention, and transforming a cellulase-producing organism with the inactivated gene. Recombination between the inactivated and genomic gene leads to inactivation of the selected genomic gene. The feasibility of this approach applied to filamentous fungal organisms has been shown in Yelton, M. M., et al., Proc. Nat. Acad. Sci., USA, (1984) 80:1470.

These and other objects and aspects of the invention will become more apparent when the following detailed description is read in conjuction with the accompanying drawings.

Brief Description of the Drawings

Figure 1 shows gel electrophoretic patterns of poly A RNA isolated from T. reesei grown either in the presence (lane 1) or absence (lane 2) of a cellulase-inducing carbon source.

Figure 2 shows gel electrophoretic patterns of polypeptides produced in a cell-free protein synthesizing system supplied with:

(a) total poly A RNA obtained from T. reesei grown under cellulase-induction conditions (lanes 1 and 2); (b) three different size fractions of the total induced poly A RNA (lanes 3-8); or (c) a size fraction of poly A RNA obtained from T. reesei cells grown under cellulase non-induction conditions (lanes 9 and 10), where the odd-numbered gel patterns show total polypeptide synthesized under the direction of the various RNA samples, and the associated even-numbered gels show the corresponding peptides which immunoprecipitate in the presence of anti-CBHI antibody.

Figure 3A is a map of a vector designated pCBH157 which contains a 1.16 kilobase HindIII fragment from T. reesei strain L27 genomic DNA.

Figure 3B is a map of a vector designated pCBH164 containing a 2.3 kilobase HindIII fragment adjacent the above 1.16 kilobase fragment from the same genomic DNA.

Figure 4 is a map of the two adjacent HindIII genomic digest fragments from T. reesei strain L27 which, together, contain the CBHI gene.

Figure 5 is a map of an E. coli/S. cerevisiae shuttle vector used in one embodiment of the invention.

Figure 6 illustrates steps in the construction of an intron-free 5'-end portion of a CBHI gene, according to one embodiment of the invention.

Figure 7 illustrates steps in the construction of a 3'-end portion of an intron-free CBHI gene, according to the embodiment of the invention.

Figure 8 illustrates the construction of a vector designated ptrpCBH5 from pDG151 (Figure 5) and pCBH5 (Figure 6).

Figure 9 illustrates the construction, from ptrpCBH5 (Figure 8) and pCBH3 (Figure 7), of a vector designated ptrpCBH8 which contains a full length, intron-free CBHI gene.

Figure 10 illustrates an alternative construction, from vectors ptrpCBH5 and pCBH3, of a vector designated ptrpCBH81 containing a full-length, intron-free CBHI gene.

Figure 11 shows the construction of ptrpCBH82 a vector suitable for expression of CBHI in E. coli.

Figure 12 shows the construction of peno500.202 a vector suitable for expression of CBHI in yeast.

Figure 13 shows the results of Western blot analysis of native and recombinant CBHI.

Figure 14 illustrates in frame A, a restriction map of the EGI-encoding portion of pUC8-h; in frames B-D, the series of steps resulting in the construction of an intron-free coding sequence for EGI; and in frames E and F, the steps in providing HindIII sites immediately upstream and downstream of the EGI coding sequence.

Figure 15 shows the results of a hybridization experiment testing the ability of a 4 kb HindIII fragment from T. reesei to isolate mRNA capable of directing EGI synthesis.

Figure 16 shows the ability of colonies transformed with pJV160 to hydrolyze CMC.

Figure 17 is an autoradiogram of Western blots performed on culture supernatants from yeast transformed with pJV160 and with a control vector.

Modes of Carrying Out the Invention

A. Definitions

"Cellobiohydrolase" or "exo-cellobiohydrolase" or "CBH", as the terms are used herein, refers to a protein capable of cleaving cellulose into cellobiose units from the non-reducing ends of cellulose polymer chains. CBH enzymes are known to occur in filamentous bacterial and fungal sources, but have yet to be demonstrated in non-filamentous bacteria. At least two forms of CBH

are known in Trichoderma -- CBHI which generates only cellobiose residues, and CBHII which generates a mixture of cellibiose and glucose.

The CBHI protein illustrated herein has a sequence of 496 amino acids and is preceded by a 17 amino acid leader sequence. The mature CBHI, produced by removal of the leader sequence, has an approximate molecular weight of 66,000 daltons. When produced in native form the protein shows some degree of glycosylation predominantly through linkage to the OH of serine and/or threonine (Gum, E.K., et al., Biochem. Biophys. Acta (1976) 446:371).

"Endoglucanase" refers to an enzyme which is capable of cleaving cellulose at internal glycosidic sites. A number of forms of endoglucanase are known including both EGI and EGII from fungal sources, such as T. reesei and other suitable fungal sources.

The EGI illustrated herein has a sequence of 437 amino acids in what is considered to be the mature protein, and is preceded by a 22 amino acid leader sequence. The secreted form of the protein, based on its 437 amino acid composition, has a molecular weight of about 46,000 daltons. The native protein is known to contain some degree of glycosylation when produced by the native host. (Shoemaker, S.P., et al., Biochem. Biophys. Acta (1978) 523:147).

"β-glucosidases" are similarly defined. These enzymes comprise part of the cellulase system of fungi, as well as of bacterial systems. They are known to exist in at least two physical forms, β-glucosidase-I, and β-glucosidase-II in T. reesei. These enzymes complete the cellulose digestion by generating glucose units from soluble oligosaccharide. The recombinant β-glucosidases of the invention include proteins with this activity and specificity whose amino acid sequences are substantially equivalent to that of a fungal β-glucosidase, whether or not absolutely identical.

The following applies to all three types of the cellulases defined above:

It is clear that all proteins are potentially either acidic or basic salts according to the pH at which they are prepared. All ionization states of the protein are included in the definition. Further, it is understood that certain modifications are made to the

5 molecule by virtue of post-translational processing, such as glycosylation, phosphorylation, acetylation, or the like, which are dependent on the nature of the host system. Included in the definitions of the cellulases are both processed and unprocessed forms of the polypeptides. Similarly, modifications to the structure may

10 occur through inadvertent or intentional chemical modification of amino acid side chains such as, for example, oxidation of sulfhydryl groups. Proteins so modified are further included in the definition so long as these modifications do not destroy enzyme activity. Finally, it is expected that minor alterations in the precise amino

15 acid sequences may be made without destroying the activity of the enzymes. Such modifications may include amino acid substitutions as well as additions or deletions of one or a few amino acids in the sequence, and these modifications, too are included in the definition, provided enzyme activity is retained.

20 In particular, two cellulase enzymes have "substantially equivalent" amino acid sequences when an approximate one-to-one correspondence between the sequences exists; however, minor changes in one or several amino acids which do not result in loss of its CBH, EG, or β-glucosidase functionality are considered to generate peptides

25 remaining within this definition.

A particular cellulase "derived from" a particular source refers to amino acid sequences which are produced natively by that source.

In the illustrations below, the entire genomic sequences

30 associated with EGI or CBHI derived from T. reesei L27 are used as the coding sequences. These coding sequences contain the pre-sequences encoding the leader peptides. In these illustrations, the pre-sequences are retained because it may be advantageous to obtain secretion of the proteins produced. However, in other circumstances,

it may be desirable to produce mature proteins which are not secreted from the cells. Means for modifying the coding sequence so as to delete this leader or portions thereof, thus effecting a production of non-secreted form of the mature protein are well within the scope of

5   the art. In addition to such mature proteins, the invention also encompasses fused forms of cellobiohydrolase, endoglucanase or β-glucosidase, which are or are not secreted, and the associated production means and materials.

The target gene will be referred to as the EGI or CBHI gene

10   in the illustrations below. It will be clear from the context what the precise nature of this gene is, i.e., whether it encodes a mature, fused, or leader sequence-containing protein, and it will be so referenced whether or not it includes a pre-sequence and regardless of its form as mature or fused. The DNA sequences of the invention are

15   those which encode active protein, regardless of any additional (or deleted) nucleotides beyond those of the mature form.

"Treatment" of a suitable cellulase substrate, such as cellulose, encompasses hydrolytic enzyme reactions designed either to produce selected substrate modifications, or to effect hydrolytic

20   breakdown of the substrate, including complete hydrolysis of cellulose to glucose, or limited, controlled hydrolysis/modification.

"Control sequences" refer to DNA sequences which regulate the expression of a particular coding sequence in a particular host. Depending on the nature of the host these may include promoters,

25   operators, ribosome binding sites, terminators, enhancers, and whatever other sequences are necessary to effect some level of expression. "Suitable" control sequences refer to those sequences which are compatible with the particular host concerned.

One of the control sequences, termed "signal sequence", is

30   defined herein as a sequence of amino acids which is responsible for initiating export of a protein chain. A signal sequence, once having initiated export of a growing protein chain, is cleaved from the mature protein at a specific site. The term also includes leader sequences or leader peptides. One preferred signal sequence herein is

the deduced signal sequence for CBHI derived from <u>T. reesei</u> indicated in Table I. Another is that for EGI from <u>T. reesei</u> shown in Table III.

"Operably linked" refers to juxtaposition of gene elements and for control sequences in such a manner that the activity of the components of the system is retained. Specifically, "coding sequences operably linked to control sequences" refers to a configuration wherein the coding sequences are expressed in the appropriate host. The control sequences provided may be maximized as to their efficiency, but the phrase "operably linked to suitable control sequences" simply implies that sufficient control sequence efficacy is available to provide some useful level of expression.

"Cell", "cell line", "cell culture", and "host cell" are used interchangeably and meant to denote the most general case unless otherwise clear from the context.

A vector which is a "derivative" of another vector means that the parent vector was employed in the series of steps leading to the synthesis of the "derivative" vector.

B. <u>General Description</u>

B.1. <u>Production of Recombinant Cellulases</u>

In general terms, the production of a recombinant form of cellulase enzyme in the approach illustrated below involves the following steps:

1) obtaining a coding sequence uninterrupted by introns for the mature cellulase protein, for a fused form of the protein, or for the protein preceded by a leader sequence in excisable or otherwise recoverable form;

2) placing the cellulase coding sequence, either before or after removal of intron(s) according to step 1, in operable linkage with suitable control sequences of a replicable expression vector;

3) transforming a suitable host with the thus formed expression vector;

4)   culturing the transformed host under favorable conditions to effect the production of the recombinant cellulase enzyme; and,

5) optionally recovering the cellulase from the medium or from the cells. Recovery of the enzyme may, in fact, not be necessary in order to utilize its activity for cellulose degradation. Indeed, in one embodiment of the method of the invention, which concerns use of the recombinant cellulase for cellulose degradation, cellulose may be added to the cultured cells in situ, and the growing or resting cell culture used to effect cellulose degradation directly.

Each of the foregoing steps can be done in a variety of ways. For example, the desired coding sequences can be obtained by preparing suitable cDNA from cellular messenger and manipulating the cDNA to obtain the complete sequence. Alternatively, genomic fragments may be manipulated, according to a number of procedures such as those described below, to remove identified intron regions. If a portion of a sequence is sufficiently short to come within the purview of known chemical synthetic techniques, such sequence can be prepared from the individual nucleotide starting materials through chemical synthesis. Indeed, the length of sequence obtainable by this method is entirely dependent on the current status of the oligonucleotide synthesis art.

A particularly useful way to retrieve the coding sequences corresponding to the cellulase enzymes from a genomic library is illustrated herein. The genomic library, prepared in phage from the nuclear DNA of a high producer for the cellulase complex, such as induced T. reesei L27, contains coding sequences for all the desired enzymes. For each enzyme desired, a purified native form is used to obtain antibodies. The antibodies are used to identify an enriched mRNA fraction capable of synthesizing the appropriately immunoprecipitating protein in an in vitro translation system. The identified mRNA is used to generate a cDNA probe which, in turn, is used to identify the desired gene or gene portions from the library. Further modifications in the probe-identified gene may be made. For

convenience, suitable restriction sites can, if not natively available, be added to the ends of the coding sequence so as to provide an excisable gene segment which can then be inserted adjacent suitable control sequences in an expression vector. Alternatively,
5   control sequences may be introduced into vectors already containing the gene or portions thereof. The control sequences, expression vectors, and transformation methods are dependent on the type of host cell used to express the gene. Generally, procaryotic, yeast, or mammalian cells are presently useful as hosts. The appropriate
10   control sequences and techniques will be dependent on the choice of host as is outlined below.

B.1.a. <u>Control Sequences</u>

Procaryotes most frequently are represented by various strains of <u>E. coli</u>. However, other microbial strains may also be
15   used, such as bacilli, for example <u>Bacillus subtilis</u>, various species of <u>Pseudomonas</u>, or other bacterial strains. In such procaryotic systems, plasmid vectors which contain replication sites and control sequences derived from a species compatible with the host are used. For example, <u>E. coli</u> is typically transformed using derivatives of
20   pBR322, a plasmid derived from an <u>E. coli</u> species by Bolivar, et al., <u>Gene</u> (1977) <u>2</u>:95. pBR322 contains genes for ampicillin and tetracycline resistance, and these markers can be either retained or destroyed in constructing the desired vector. Commonly used procaryotic control sequences which are defined herein to include
25   promoters for transcription initiation, optionally with an operator, along with ribosome binding site sequences, include such commonly used promoters as the beta-lactamase (penicillinase) and lactose (lac) promoter systems (Chang, et al., <u>Nature</u> (1977) <u>198</u>:1056 and the tryptophan (trp) promoter system (Goeddel, et al., <u>Nucleic Acids Res</u>.
30   (1980) <u>8</u>:4057. The lambda-derived P_L promoter and N-gene ribosome binding site (Shimatake, et al., <u>Nature</u> (1981) <u>292</u>:128), which has been made useful as a portable control cassette, as set forth in copending Application Serial No. 578,133, filed February 8, 1984, and

assigned to the same assignee, is another example. However, any available promoter system compatible with procaryotes can be used.

In addition to bacteria, eucaryotic microbes, such as yeast, may also be used as hosts. Laboratory strains of _Saccharomyces_
5 _cerevisiae_, Baker's yeast, are most used, although a number of other strains are commonly available. While vectors employing the 2 micron origin of replication are illustrated, (Broach, J. R., _Meth. Enz._ (1983) _101_:307), other plasmid vectors suitable for yeast expression are known (see, for example, Stinchcomb, et al., _Nature_ (1979) _282_:39,
10 Tschempe, et al., _Gene_ (1980) _10_:157 and Clarke, L., et al., _Meth._ _Enz._ (1983) _101_:300). Control sequences for yeast vectors include promoters for the synthesis of glycolytic enzymes (Hess, et al., _J._ _Adv. Enzyme Reg._ (1968) _7_:149; Holland, et al., _Biochemistry_ (1978) _17_:4900). Additional promoters known in the art include the promoter
15 for 3-phosphoglycerate kinase (Hitzeman, et al., _J. Biol. Chem._ (1980) _255_:2073), and those for other glycolytic enzymes, such as glyceraldehyde-3-phosphate dehydrogenase, hexokinase, pyruvate decarboxylase, phosphofructokinase, glucose-6-phosphate isomerase, 3-phosphoglycerate mutase, pyruvate kinase, triosephosphate isomerase,
20 phosphoglucose isomerase, and glucokinase. Other promoters, which have the additional advantage of transcription controlled by growth conditions are the promoter regions for alcohol dehydrogenase 2, isocytochrome C, acid phosphatase, degradative enzymes associated with nitrogen metabolism, and enzymes responsible for maltose and galactose
25 utilization (Holland, ibid).

Evidence suggests that terminator sequences are desirable at the 3' end of the coding sequences. Such terminators are found in the 3' untranslated region following the coding sequences in yeast-derived genes. Many of the vectors illustrated contain control sequences
30 derived from the enolase-I gene containing plasmid peno46 (Holland, M. J., et al., _J. Biol. Chem._ (1981) _256_:1385) or the LEU2 gene obtained from YEp13 (Broach, J., et al., _Gene_ (1979) _8_:121), however any vector containing a yeast compatible promoter, origin of replication and other control sequences is suitable.

It is also, of course, possible to express genes encoding polypeptides in eucaryotic host cell cultures derived from multicellular organisms. See, for example, Tissue Cultures, Academic Press, Cruz and Patterson, editors (1973). Useful host cell lines

5 include VERO and HeLa cells, and Chinese hamster ovary (CHO) cells. Expression vectors for such cells ordinarily include promoters and control sequences compatible with mammalian cells such as, for example, the commonly used early and late promoters from Simian Virus 40 (SV 40) Fiers, et al., Nature (1978) 273:113), or other viral

10 promoters such as those derived from polyoma, Adenovirus 2, bovine papiloma virus, or avian sarcoma viruses. General aspects of mammalian cell host system transformations have been described by Axel in U.S. Patent No. 4,399,216 issued August 16, 1983. It now appears also that "enhancer" regions are important in optimizing expression;

15 these are, generally, sequences found upstream of the promoter region. Origins of replication may be obtained, if needed, from viral sources. However, gene integration into the chromosome is a common mechanism for DNA replication in eucaryotes, and hence independently replicating vectors are not required. Plant cells are also now

20 available as hosts, and control sequences compatible with plant cells such as the nopaline synthase promoter and polyadenylation signal sequences (Depicker, A., et al., J. Mol. Appl. Gen. (1982) 1:561) are available.

B.1.b. Transformations

25 Depending on the host cell used, transformation is done using standard techniques appropriate to such cells. The calcium treatment employing calcium chloride, as described by Cohen, S. N., Proc. Natl. Acad. Sci. (USA) (1972) 69:2110 is used for procaryotes or other cells which contain substantial cell wall barriers. Infection

30 with Agrobacterium tumefaciens (Shaw, C. H., et al., Gene (1983) 23:315) is used for certain plant cells. For mammalian cells without such cell walls, the calcium phosphate precipitation method of Graham and van der Eb, Virology (1978) 52:546 is preferred. Transformations into yeast were carried out according to the method of Van Solingen,

P., et al., J. Bact. (1977) 130:946 and Hsiao, C. L., et al., Proc. Natl. Acad. Sci. (USA) (1979) 76:3829. Briefly, yeast cultures grown to mid-log phase in YEPD rich medium (yeast extract, peptone and 4% glucose) were washed and protoplasted with zymolyase 5000 (Miles Laboratory) in sorbitol phosphate buffer. Protoplasts were washed, allowed to stand at room temperature for one hour in 67% YEPD containing 1 M sorbitol, then pelleted and suspended in Tris-sorbitol-calcium chloride buffer to 2×10^9 protoplasts/ml. Protoplasts were mixed with 5-10 µg of DNA for transformation in a 100 µl reaction mix, then 1 ml of 44% PEG was added and the mixture allowed to stand for 40 minutes at room temperature. Alternatively, the procedure of Klebe, et al., (Gene (1983) 25:333), can be used.

B.1.c. Vector Construction

Construction of suitable vectors containing the desired coding and control sequences employs standard ligation and restriction techniques which are well understood in the art. Isolated plasmids, DNA sequences, or synthesized oligodeoxyribonucleotides are cleaved, tailored, and religated in the form desired.

Site specific DNA cleavage is performed by treating with the suitable restriction enzyme (or enzymes) under conditions which are generally understood in the art, and the particulars of which are specified by the manufacturer of these commercially available restriction enzymes. See, e.g., New England Biolabs, Product Catalog. In general, about 1 µg of plasmid or DNA sequence is cleaved by one unit of enzyme in about 20 µl of buffer solution; in the examples herein, typically, an excess of restriction enzyme is used to insure complete digestion of the DNA substrate. Incubation times of about one hour to two hours at about 37°C are workable, although variations can be tolerated. After each incubation, protein is removed by extraction with phenol/chloroform, and may be followed by ether extraction, and the nucleic acid recovered from aqueous fractions by precipitation with ethanol followed by running over a Sephadex G-50 spin column. If desired, size separation of the cleaved fragments may be performed by polyacrylamide gel or agarose gel

electrophoresis using standard techniques. A general description of size separations is found in Methods in Enzymology (1980) 65:499-560.

Restriction cleaved fragments may be blunt ended by treating with the large fragment of E. coli DNA polymerase I (Klenow) in the presence of the four deoxynucleotide triphosphates (dNTPs) using incubation times of about 15 to 25 min. at 20 to 25°C in 50 mM Tris-Cl, pH 7.6, 50 mM NaCl, 6 mM $MgCl_2$, 6 mM DTT and 5-10 μM dNTPs. The Klenow fragment fills in at 5' sticky ends but chews back protruding 3' single strands, even though the four dNTPs are present. If desired, selective repair can be performed by supplying only a selected one or more dNTPs within the limitations dictated by the nature of the sticky ends. After treatment with Klenow, the mixture is extracted with phenol/chloroform and ethanol precipitated followed by running over a Sephadex G-50 spin column. Treatment under appropriate conditions with S1 nuclease results in hydrolysis of any single-stranded portion.

Synthetic oligonucleotides may be prepared by the triester method of Matteucci, et al. (J. Am. Chem. Soc. (1981) 103:3185-3191), or the diethylphosphoramidite method of Caruthers, described in U.S. Patent No. 4,415,732, issued November 15, 1983.

Kinasing of single strands prior to annealing or for radio-labeling is achieved using an excess polynucleotide kinase, e.g., approximately 10 units of kinase to 1 nmole substrate in the presence of 50 mM Tris, pH 7.6, 10 mM $MgCl_2$, 5 mM dithiothreitol, 1-2 mM ATP, 1.7 pmoles γ32P-ATP (2.9 mCi/mmole), 0.1 mM spermidine, and 0.1 mM EDTA.

Ligations are performed typically in 15-30 μl volumes under the following standard conditions and temperatures: 20 mM Tris-Cl pH 7.5, 10 mM $MgCl_2$, 10 mM DTT, 33 μg/ml BSA, 10 mM-50 mM NaCl, and either 40 μM ATP, 0.01-0.02 (Weiss) units T4 DNA ligase at 0°C (for "sticky end" ligation) or 1 mM ATP, 0.3-0.6 (Weiss) units T4 DNA ligase at 14°C (for "blunt end" ligation). Intermolecular "sticky end" ligations are usually performed at 33-100 μg/ml total DNA concentrations (5-100 nM total end concentration). Intermolecular

blunt end ligations (usually employing a 10-30 fold molar excess of linkers) are performed typically at about 1 μM total ends concentration.

In vector constructions employing "vector fragments", the vector fragment is commonly treated with bacterial alkaline phosphatase (BAP) in order to remove the 5' phosphate and prevent religation of the vector. BAP digestions are conducted typically at about pH 8, in approximately 150 mM Tris, in the presence of Na^+ and Mg^{+2}, using about 1 unit of BAP per μg of vector at 60° for about one hour. In order to recover the nucleic acid fragments, the preparation is extracted with phenol/chloroform and ethanol precipitated and desalted by application to a Sephadex G-50 spin column. Alternatively, religation can be prevented in vectors which have been double digested by additional restriction enzyme digestion of the unwanted fragments.

B.1.d. Verification of Construction

In the constructions set forth below, correct ligations for plasmid construction are confirmed by first transforming E. coli strain MM294 obtained from E. coli Genetic Stock Center, CGSC #6135, or other suitable host with the ligation mixture. Successful transformants are selected by ampicillin, tetracycline or other antibiotic resistance or using other markers depending on the mode of plasmid construction, as is understood in the art. Plasmids from the transformants are then prepared according to the method of Clewell, D. B., et al., Proc. Natl. Acad. Sci. (USA) (1969) 62:1159, optionally following chloramphenicol amplification (Clewell, D. B., J. Bacteriol. (1972) 110:667). The isolated DNA is analyzed by restriction and/or sequenced by the dideoxy method of Sanger, F. et al., Proc. Natl. Acad. Sci. (USA) (1977) 74:5463 as further described by Messing, et al., Nucleic Acids Res. (1981) 9:309, or by the method of Maxam, et al., Methods in Enzymology (1980) 65:499.

B.1.e. <u>Site Specific Mutagenesis</u>

Oligonucleotide induced mutagenesis is conducted using a primer synthetic oligonucleotide complementary to a single stranded phage DNA to be mutagenized except for limited mismatching, representing the desired mutation. Briefly, the synthetic oligonucleotide is used as a primer to direct synthesis of a strand complementary to the phage, and the resulting double-stranded DNA is transformed into a phage-supporting host bacterium. Cultures of the transformed bacteria are plated in top agar, permitting plaque formation from single cells which harbor the phage.

Theoretically, 50% of the new colonies will contain the phage having, as a single strand, the mutated form; 50% will have the original sequence. The resulting plaques are hybridized with kinased synthetic primer at a temperature which permits hybridization of an exact match, but at which the mismatches with the original strand are sufficient to prevent hybridization. Plaques which hybridize with the probe are then picked, cultured, and the DNA recovered. Details of site specific mutation procedures are described below in specific examples.

B.1.f. <u>Hosts Exemplified</u>

Host strains used in cloning and expression herein are as follows:

For cloning and sequencing, and for expression of construction under control of most bacterial promoters, <u>E. coli</u> strain MM294 (supra), Talmadge, K., et al., <u>Gene</u> (1980) <u>12</u>:235; Meselson, M., et al., <u>Nature</u> (1968) <u>217</u>:1110, was used as the host.

For M13 phage recombinants, <u>E. coli</u> strains susceptible to phage infection, such as <u>E. coli</u> K12 strain DG98 are employed. The DG98 strain has been deposited with ATTC and has accession number <u>39768</u> .

Expression in yeast employed laboratory strains of <u>S. cerevisiae</u>: C468 (α, his4$^-$, leu2$^-$, mal$^-$) strain obtainable from ATCC,

Deposit 20,690, and S173-6B (trp1⁻, leu2⁻, ura3⁻, his4⁻) strain, obtainable from Prof. Fred Sherman, U. of Rochester, N.Y.

B.2. Utility of the Recombinant Enzyme

Recombinant cellobiohydrolases, endoglucanases, and β-glucosidases, can, of course, be employed for approximately the same utilities, and under the same conditions as are appropriate for the native forms of the corresponding enzyme. The recombinant forms have the advantage of being obtainable singly or in controlled combination in relatively large quantity, and of being subject to suitable modification to enhance their properties by genetic manipulation of the coding sequences or by selecting hosts which have optimal post-translational processing features The recombinant enzymes, when produced in a novel host, confer in the host the ability to utilize and/or modify novel substrates.

Each of the recombinant cellulase enzymes can be employed in vitro, either in solution or in immobilized form, to carry out its role in the complete or controlled hydrolysis of cellulose, or such other modified substrates as is desired and compatible with the enzyme in question. The specific conditions of temperature, pH, ionic strength, and time of reaction will be variable with the precise enzymes employed.

The enzymes of the invention can be used either alone or in combination with each other or with other suitable catalysts. For efficient hydrolysis of an insoluble cellulose substrate, it is clearly advantageous to use a combination of at least one endoglucanase, cellobiohydrolyase, and β-glucosidase. Again, these can be supplied in soluble or immobilized form, or can be provided by the same or different transformant host. For example, it is within the scope of the invention to transform a yeast host with expression vectors which contain each of the enzymes of the cellulase complex, or with a single expression vector which contains expression modules for each of these enzymes. It is currently known that some species of yeast endogenously produce β-glucosidase and other strains may be found which endogenously produce other enzymes of the cellulase

found which endogenously produce other enzymes of the cellulase complex. For such host strains it would be necessary only to supply the missing member(s) of the cellulase complex in the form of suitable expression vector(s).

5   Further, the utility of the recombinant enzymes of the invention is not limited to the production of glucose from cellulose. It has already been mentioned that the specificity of EGII may be advantageously employed to facilitate the preparation of freeze-dried coffee by effecting the hydolysis of an interferring
10   polysaccharide in coffee beans. The CBH's are valuable research tools because of their high specificity for 1,4-β,D-glucan linkages, and thus provide a means for ascertaining the presence or absence of such linkages in preparation of polysaccharides. Indeed, under some circumstances it might be desirable to obtain a pure preparation of
15   cellibiose specifically, and CBHI would be singularly useful for this purpose. A method for engineering a cellulase-producing organism to produce only CBHI cellobiohydrolase has been outlined above. Finally, industrial yeast strains may be transformed with one or more of the components of the cellulase complex simply by providing the vector
20   with a suitable dominant selectable marker. See, The Molecular Biology of Yeast, Cold Spring Harbor Meeting, August 16-21, 1983.

B.3   Recombinant-Gene Yeast Vectors

          As will be seen below, recombinant vectors constructed according to the present invention provide efficient production in
25   yeast of secreted, glycosylated, biologically active cellulases. The novel constructional features of the vector which are important in yeast to production of heterologous recombinant cellulases include: (1) a yeast promoter sequence; (2) a cellulase enzyme signal sequence linked to the promoter; and (3) the coding region of the mature
30   cellulase to be produced. Additionally, the exemplary recombinant vectors contain a yeast terminator sequence at the 3'-end of the cellulase coding region.

According to one aspect of the invention, the yeast vectors containing cellulase genes may be modified, by conventional recombinant techniques, to include: (1) the above yeast promoter; (2) the cellulase enzyme signal sequence; and (3) a suitable restriction site at the 3' end of the signal sequence or within the adjacent

5 cellulase coding region at which a heterologous gene can be introduced. The heterologous gene may code for another cellulase, or for an unrelated protein, such as an amylase, protease, peptide hormone, or one of a large number of other peptides for which a heterologous gene, preferably in uninterrupted form, can be obtained.

10 Section D below illustrates an exemplary yeast vector designated penoCBH500.202, containing a yeast enolase promoter, a CBHI signal sequence, the coding sequence of mature CBHI from T. reesei, and the yeast enolase terminator. The vector can be modified according to known methods, to replace all or a portion of the CBHI

15 coding sequence with a heterologous protein coding sequence. The resultant vector allows the efficient processing and secretion of the heterologous protein in yeast.

Section E below describes another exemplary yeast vector, designated pJV160, containing a yeast enolase promoter, an EGI signal

20 sequence, the coding sequence of mature EGI from T. reesei, and the above enolase terminator sequence. Using analogous techniques, this vector can be modified to provide efficient processing and secretion of a heterologous-gene product in yeast.

C. Cloning and Expression of EGI and CBHI from T. reesei Strain L27

25 Illustrated below in Sections D and E are embodiments of the present invention wherein the DNA sequences encoding the CBHI and EGI secreted from T. reesei strain L27, respectively, are cloned and expressed. These illustrative embodiments are not intended to limit the invention, but rather to describe exemplary workable approaches to

30 obtaining recombinant cellulase enzymes generally, and expression vectors permitting production of glycosylated, mature, enzymatically active fungal cellulase enzymes in a heterologous host such as yeast.

Briefly, and in summary, a preferred source of the two coding sequences is a strain of T. reesei, strain L27, deposited in the NRRL Culture Depository and given accession number NRRL #15538. This strain was selected as a high level producer of the cellulase

5 complex and because portions of the EGI and CBHI sequences derived from T. reesei have been published (referenced below).

The strategy for obtaining the intronless coding sequences was as follows: poly-A RNA (mRNA) was isolated from the selected fungal source following cellulose induction, and size-fractionated

10 using agarose gel electrophoresis. Gels containing mRNA fractions derived from induced cells, (which showed a number of bands absent or present only in relatively small amounts in uninduced cells) were sliced according to size, and each slice tested for the presence of the desired mRNA. Testing was conducted by adding eluted mRNA to an

15 in vitro protein synthesis (translation) system, and analyzing the proteins synthesized by SDS-PAGE both by electrophoresis of total proteins and polypeptides which immunoprecipitate in the presence of anti-EGI or anti-CBHI antibodies. This procedure enabled identification in each case of an mRNA fraction enriched in the

20 desired message.

The identified mRNA fraction was then used to obtain single stranded copy DNA (cDNA), which was radiolabelled and used as a probe of a genomic library. The cDNA was also used to obtain a cDNA library which included the fragment(s) containing coding sequences for the EGI

25 or CBHI gene.

A genomic DNA library from the selected fungal strain was prepared in a λ phage. The total nuclear DNA from the fungus was partially digested with one or more selected restriction enzymes, and inserted in the phage vector. This phage library contains coding

30 sequences for all the enzymes of this fungal cellulase complex. The cloned fragments were probed with cDNA derived from the appropriate enriched mRNA fraction. Phages homologous to the CBHI or EGI specific probes were identified and the T. reesei DNA fragments corresponding to the coding regions were subcloned into suitable plasmid vectors. A

4 kb HindIII genomic fragment containing the entire EGI sequence, and two contiguous HindIII genomic fragments of 1.16 kb and 2.3 kb which together encode the entire CBHI sequence were obtained by this method.

5   By further restriction analysis and sequencing, the sequence of the coding portion of the selected fragments were determined. Gene sequencing led to the identification of two introns in both CBHI and EGI genes. The cellulase gene introns are removed, preferably before the gene is placed in its expression vector.

10   For CBHI, the two introns were removed by cDNA fragment replacement. For EGI, one intron was deleted by replacing a portion of the genomic DNA with the corresponding sequences of a cDNA clone obtained from the cDNA library; the other intron was deleted by site directed mutagenesis using a chemically synthesized primer.

15   The intronless CBHI coding fragments were used to construct an expression vector capable of transforming yeast to secrete functionally active CBHI. For EGI, the ends of the coding sequences were modified by site directed mutagenesis to insert desirable restriction sites, making the entire coding sequence available as a HindIII cassete. This cassette was cloned into an expression vector operable in yeast, and used to transform a suitable yeast host, resulting in a recombinant cell capable of EGI production.

20   In analogous fashion, the intronless coding sequences for EGII, CBHII, and β-glucosidase-I and β-glucosidase-II are obtained, introduced into suitable host vectors and transformed into a desired host, such as yeast. Antibody preparations have, indeed, been prepared to β-glucosidase-II in a manner similar to that illustrated below, the correct mRNA fraction identified, and a cDNA probe (single stranded) and a cloning vector cDNA library (double stranded) prepared for this enzyme.

25

30   D. Preparation of Recombinant CBHI

This section describes an intron-free cellobiohydrolase-I (CBHI) gene, expression vectors into which the gene has been introduced by recombinant DNA techniques, hosts transformed with these

vectors and CBHI prepared thereby. The CBHI gene is derived from a selected fungal source which preferably is capable of producing induced levels of extracellular cellulases, including CBHI, when grown on an inducing carbon source, such as cellulose.

5   D.1 Selection of a Fungal Source of CBHI

In selecting a fungal source of CBHI for use in the present invention, it is advantageous to select an organism in which CBHI and the corresponding mRNA are inducible from substantially undetectable activity levels, when grown on a non-inducing carbon source, to easily measured levels when grown on an inducing carbon source, such as cellulose. Preferred fungal sources includes species of the Trichoderma genus, and most preferably T. reesei, which has an inducible CBHI enzyme that constitutes up to about 60% of the total extracellular cellulases induced when the organism is grown in the presence of cellulose. An additional advantage of T. reesei is that the enzymes which make up the inducible cellulase system have been relatively well studied. In particular, a substantial portion of the amino acid sequence of the T. reesei CBHI enzyme has been reported (Fagerstam, L. et al., FEBS LETT. (1980) 119:97 and Fagerstam, L. G., Doctoral Thesis, University of Uppsala, Sweden (1981)).

The ability of the microbial source to produce CBHI mRNA efficiently and in large amounts may be further enhanced by colony selection techniques, preferably after exposing the microbial source to a mutagen, to identify strains which are capable of producing elevated levels of one or more cellulases and/or which are capable of producing maximally induced levels of cellulase over reduced fermentation periods. One of the inventors herein (Shoemaker) has previously described a group of T. reesei mutant strains capable of producing higher induced levels of cellulase enzymes than in parent, commercially available strains. Further, as noted below, the maximal increase in extracellular cellulases was observed after a shorter enzyme-induction period than in the parent strain. The reader is referred to Shoemaker, S. P., et al., Trends in the Biology of Fermentations for Fuels and Chemicals, Hollaender et al. (eds), Plenum

Publishing Corp., N.Y., N.Y., (1981) p. 89 for a detailed discussion of methods for obtaining mutant T. reesei strains of this type. What will be described herein are comparative properties of one of the mutant strains, designated strains L27, and the parent strain,
5   Quartermaster No. 9414 (QM 9414), which led to the selection of the L27 strain as a preferred candidate for use in the invention. T. reesei strain L27 is deposited in the NRRL culture depository and characterized by NRRL #15538.

The higher levels of extracellular cellulases produced by
10   strain L27 suggest that this strain may produce higher levels of cellulase mRNA's than does strain QM 9414. To determine relative cellulase mRNA levels in the two strains, total poly A RNA from each of the two strains grown under cellulase-inducing conditions was isolated and added to a cell-free protein synthesizing system to
15   direct the synthesis of cellulase polypeptides. General procedures used in performing the cell-free protein syntheses are described in paragraph D.3 below. The polypeptides synthesized in each cell-free synthesizing experiment were immunoprecipitated with antibodies to either CBHI, EGI, or beta-glucosidase. The immunoprecipitates were
20   examined conventionally by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), referenced in subparagraph C.3. As judged by the relative amounts of radioactivity in specific protein-band regions of the gel (not shown), strain L27 gave higher in vitro translation activity for all three enzymes, suggesting that the strain
25   produces higher in vivo levels of mRNA for each cellulase component studied.

In the above experiment, the total poly A RNA was isolated from the induced cultures at the time when greatest increase in extracellular cellulases was observed. This induction time was
30   between about 18 to 20 hours for strain L27, and between about 26 and 28 hours for strain QM 9414. Therefore, strain L27 provides the additional advantage that relatively shorter induction times are required for the production of induced cellulase mRNA's. It is also noted here that strain L27 shows induced cellulase levels when grown
35   on either cellulose or glucose, unlike the parent strain QM 9414 in

which cellulase synthesis is glucose-repressed.

D.2 Purification of CBHI Protein and Production of Anti CBHI Antibodies

According to one feature of the invention, CBHI mRNA may be
5   identified by its ability to direct the synthesis, in a cell-free protein synthesizing system, of a polypeptide which is immunoreactive with anti-CBHI antibody. The antibody is preferably raised against a purified CBHI enzyme obtained from the selected fungal source of CBHI mRNA.

10 Purified CBHI may be obtained according to standard protein purification procedures from an extracellular culture broth. Typically, the broth, or filtrate, is fractionated by one or more ion-exchange chromatographic steps, and where necessary, a molecular-sieve chromatographic step, until enzyme purity is achieved. The enzyme can
·15 be assayed during purification steps by a conventional cellulase assay method, such as will be described below, and by SDS-PAGE.

By way of example, CBHI from T. reesei strain L27 was purified from an 8-day fermentation of the strain grown on 8% Avicel (microcrystalline cellulose) obtained from FMC Corp. (Rockfield,
20 MD). The fermentation broth was filtered through a 0.45 micron micropore filter to remove cellular material, then dialyzed and concentrated in 50 mM sodium acetate buffer, pH 5.5, using an Amicon Ultrafiltration Cell (Lexington, MA). The concentrated extracellular filtrate was fractionated by anion exchange chromatography, using
25 DEAE-Sepharose obtained from Pharmacia Fine Chemicals (Uppsala, Sweden). The column material was eluted with 0-300 mM sodium chloride gradient in 50 mM sodium acetate buffer, pH 5.0. CBHI eluted from the column at about 200 mM NaCl.

The enzyme was further purified by ion exchange
30 chromatography, using DEAE-Sepharose in a 50 mM sodium acetate buffer, pH 6.0. The CBHI eluted from the column also at about 200 mM NaCl in a 0-300 mM sodium chloride gradient. The presence of CBHI in the column elution fractions was identified by the presence of reducing

sugar after incubating the fractions in a solution containing 1% phosphoric acid swollen cellulose (PSC) in a 50 mM sodium acetate buffer, pH 4.5 at 45°C for thirty minutes. The identity of the reducing sugar as cellubiose was confirmed by high performance liquid chromatography (HPLC). The reader is referred to Shoemaker, S. P., et al., Biochem. Biophys. Acta (1978) 523:147 for details of exo-cellobiohydrolase assays.

The purity of the twice-chromatographed CBHI was analyzed by SDS-PAGE, using a modified discontinuous buffer system (Laemmli, U. K., et al., J. Mol. Biol. (1973) 80: 575) in a vertical slab apparatus. The electrophoretically separated proteins were detected by staining with Coomassie Blue R250. The gel showed a single stained band, indicating substantial enzyme purity. The apparent protein molecular weight was estimated, by comparing its migration rate with that of known molecular-weight proteins, to be about 66,000 daltons.

A comparison was made between the amino acid composition and sequence of the purified L27 CBHI and the amino acid composition and sequence of the CBHI enzyme obtained from T. reesei, QM 9414, which is disclosed in Fagerstam, L., et al. supra and Fagerstam, L. G. supra.

The amino acid compositions of the two enzymes were compared as follows: Purified CBHI from strain L27 was reduced with dithiothreitol (0.02 mM) in 6 mM guanidine and carboxymethylated with iodo[^{14}C] acetic acid (0.03 M), as described in Watt, K. W., et al., Proc. Nat. Acad. Sci. USA (1978) 75: 1731. Amino acid analysis of reduced, carboxymethylated CBHI was performed by hydrolyzing 40 μg of the protein with 0.1% phenol in 5.7 N HCl at 108° for 24, 48 and 72 hours in vacuo. The amino acids in the hydrolysate were quantified with a Spinco model 121 MB Amino Acid Analyzer (Beckman Instruments, Inc., Palo Alto, CA). Tryptophan was determined from a separate acid hydrolysate with 10% mercaptoacetic acid in 5.7 N HCl. The amino acid composition obtained for strain L27 CBHI was nearly identical to that of the published amino acid composition of CBHI from T. reesei, QM 9414.

The N-terminal portion of the strain L27 enzyme, containing the first 37 N-terminal amino acids in the mature CBHI enzyme, was sequenced, and the amino acid sequence compared with the corresponding region of the already-sequenced CBHI protein from T. reesei QM 9414.
Automated amino acid sequencing of 40-50 nmole of purified strain L27 CBHI was conducted with a spinning cup sequenator (Model 890c modified with a cold trap, Beckman Instruments, Inc., Palo Alto, CA). T. reesei strain L27 CBHI is blocked at the amino terminus with a pyroglutamyl residue. The residue is removed by treating the protein with calf liver pyroglutamate amino peptidase (Boehringer Biochemicals), as described in Podell, D. N., et al., Biochem. Biophys. Res. Commun. (1978) 81: 176. After peptidase treatment, the enzyme was filtered, lyophilized, reduced with dithiothreitol, and carboxymethylated with iodo[^{14}C] acetic acid and used directly for sequence analysis. Temporary reduction of background in sequence analysis was achieved by blocking all newly generated amino termini when a proline residue was determined to be the amino terminus. Identification of the degradation products from the sequenator was achieved on the amino acid analyzer in a fully automated HPLC system (Waters Associates).

The partial sequence as determined by this method is identical in 36 of the 37 amino acids to the published sequence of the first 37 amino acids of mature CBHI from T. reesei, QM 9414, confirming, along with the amino acid composition data, that the strain L27 CBHI is identical or nearly identical to the QM 9414 CBHI.

The purified protein is used to raise anti-CBHI antibodies by conventional methods. The antibody obtained is tested for its ability to neutralize exo-cellobiohydrolase enzyme activity, and for its ability to immunoprecipitate CBHI specifically from a mixture of induced cellulase enzymes.

To prepare antibodies against purified CBHI from T. reesei strain L27, the purified enzyme was mixed with Freunds adjuvant, and 0.2-0.4 ml of the mixture was injected intramuscularly into New Zealand white rabbits. Injections were repeated every ten days until

antibody titers were detectable by the double immunodiffusion method of Ouchterlony, O., _Arkiv Kemi._ (1949) 1:43.

The anti-CBHI antibody produced was capable of neutralizing the exo-cellobiohydrolase enzyme activity of purified CBHI.
5 Ouchterlony immunodiffusion analysis showed that the antibody prepared against the above CBHI preparation was immunoreactive against both CBHI and EGI enzymes.

D.3 Preparation of mRNA Enriched in CBHI mRNA

A preferred fungal source of CBHI, such as _T. reesei_ strain
10 L27, when grown under cellulase-induction conditions, contains a CBHI mRNA which is essentially undetectable in uninduced culture. The mRNA is capable of directing the synthesis, in a cell-free protein synthesizing system, of a polypeptide which is immunologically reactive with antibodies prepared against the CBHI protein purified
15 from the same microbial source. To obtain an RNA fraction enriched for CBHI mRNA, total poly A RNA isolated from the induced culture is size fractionated, for example, by gel-electrophoresis. Poly A RNA from induced and uninduced cultures of the microbial source may be fractionated in parallel, allowing a comparison of the gel-
20 electrophoretic patterns of the two poly A RNA preparations.

In order to identify the mRNA fraction most enriched in CBHI mRNA, various size fractions of the fractionated poly A RNA are examined for their ability to direct the synthesis, in a cell-free synthesizing system, of a polypeptide immunoprecipitable with anti-
25 CBHI antiserum. By way of illustration, procedures for preparing an mRNA fraction enriched to CBHI mRNA from _T. reesei_ strain L27 will now be described.

Mid-log _T. reesei_ strain L27 cultures grown on either Avicel (for 18 hours) or glycerol (for 48 hours) were harvested and frozen in
30 liquid nitrogen. Cultures grown on glycerol showed no measurable extracellular CBHI activity during the 48 hour growth period. Total cellular RNA was isolated from these cultures using a modification of the method described in Chirgwin, J. M., et al., _Biochem._ (1979) 18:

5294. Briefly, frozen cells were homogenized in 4.7 M guanidine thiocyanate and 12 mM adenosine:$VOSO_4$ complex (an RNase inhibitor). The RNA was pelleted through a CsCl cushion by ultracentrifugation, and the poly A RNA was isolated by oligo dT chromatography. The poly A RNA's from the induced and uninduced cultures were fractionated on methyl mercury agarose gels, according to known procedures, using standard Sea Kem agarose obtained from Marine Colloids (New York, NY). Figure 1 shows the electrophoretic patterns of poly A RNA derived from T. reesei strain L27 grown on Avicel (lane 1) or on glycerol (lane 2). Comparison of the two gels shows several RNA bands which are unique to or present in higher concentrations in induced cultures. The RNA size region corresponding to 23S RNA is shown at the left in the figure.

The gel containing induced RNA was sliced at 2 mm intervals starting at the 23S RNA region of the gel and proceeding toward lower molecular weight RNA's (toward the bottom of the gel in the figure). To release RNA from the gel slices, each slice was melted at 95°C for 5 minutes in a 0.01 M Tris-HCl buffer, pH 7.5, containing 0.1 M NaCl, 0.01 M EDTA, and 0.2% SDS. The agarose was pelleted by centrifugation. After phenol/chloroform extraction of the supernatant, the RNA was precipitated with ethanol and redissolved in water.

A small amount of each RNA fraction obtained from the gel slices was added to a cell-free protein synthesizing system, specifically a reticulocyte lysate kit obtained from New England Nuclear Company (Boston, MA) and sold as reticulocyte lysate/methionine L-[^{35}S]-translation system. Pelham, R. B., et al., Europ. J. Biochem. (1979) 67:247 describes a typical reticulocyte lysate system. After a defined reaction period, aliquots of the lysate were removed and analyzed by SDS-PAGE. The lysate material analyzed was either total lysate or the immunoprecipitate produced by adding anti-CBHI antibody to the lysate. The immunoreactive products were precipitated essentially according to conventional methods.

Selected gel electrophoretic patterns obtained are shown in Figure 2. The molecular weights shown in the figure, in thousands of daltons, are indicated by the positions of marker proteins in the lane at the left in the figure. The electrophoretic pattern of total lysate proteins produced by supplying the translation system with total "induced" poly A RNA is shown in lane 1. Lane 2 shows an analogous pattern using a hydrolyzed immunoprecipitate of these peptides with anti-CBHI. As seen, the lane 2 pattern consists of a single, relatively broad band in the molecular weight range of about 66,000 daltons, the estimated apparent molecular weight of CBHI.

As indicated above, antibody prepared against T. reesei CBHI cross reacts with T. reesei EGI, raising the possibility that at least a portion of the broad 66,000 molecular weight band seen in lane 2 is due to endogluconase polypeptide synthesis. The fact that purified EGI, when analyzed by SDS-PAGE, gives a protein band which is distinct from and has a lower molecular weight than CBHI, argues against this possibility. Also, it is expected that CBHI polypeptide synthesis would account for a large portion of the total peptide synthesis occurring in the translation system, since the CBHI enzyme makes up about 60% of the total cellulases induced in the fungal culture at the time the poly A RNA was isolated. By contrast, EGI makes up only about 5% of the total cellulases.

The size fractionated RNA obtained from successive gel slices was used to direct the synthesis of radioactive polypeptides in the translation system. Gel slice #9 --the slice located approximately 18 millimeters from the point of initial cutting of the gel at a 23S RNA band-- produced the total translation system-produced polypeptide pattern seen in lane 3, having a concentration of proteins in the 46,000-69,000 dalton molecular weight range. The analogous pattern given by the immunoprecipitate produced by adding anti-CBHI antibody is seen in lane 4, which shows a single broad protein band in the CBHI molecular weight range.

The RNA from gel slice #10, containing the next smaller size range of fractionated poly A RNA, directed the synthesis, in the

translation system, of the polypeptides seen in lane 5. Similar analysis at a hydrolyzate of the immunoprecipitate is shown in lane 6. The relatively high density of immunoprecipitated material in the CBHI molecular weight range indicates that gel slice #10 is more
5 highly enriched for CBHI mRNA than is gel slice #9.

The analogous polypeptide patterns produced by RNA from gel slice #11 are shown in lanes 7 and 8, lane 7 showing the total lysate polypeptides, and lane 8, polypeptides which immunoprecipitate in the presence of anti-CBHI antibody. As seen, the density of the
10 66,000 dalton molecular weight band in lane 8 is comparable to that in lane 4, and appreciably less than that seen in lane 6. From this comparison, gel slice #10 was identified as the gel slice maximally enriched in CBHI mRNA, and was therefore used as the source of CBHI mRNA in the procedures described below for producing a CBHI probe and
15 CBHI cDNA's.

To confirm that the relatively broad band in the 66,000 dalton molecular weight weight range seen in lane 6 represents the product of induced CBHI mRNA, the translation system was primed with a comparable-size mRNA fraction of poly A RNA from an uninduced
20 <u>T. reesei</u> culture (grown in the presence of glycerol). After fractionating the uninduced poly A RNA by gel electrophoresis, the RNA which eluted from the 10th gel slice down from the 23S region of the gel was added to the translation system. Total translation polypeptides and immunoprecipitated peptides from the translation
25 system are seen in lanes 9 and 10, respectively. As seen, lane 9 contains very little newly synthesized protein in the 66,000 molecular weight range, and lane 10 shows essentially no newly synthesized protein which immunoprecipitated in the presence of anti-CBHI antibody. The results confirm that most, if not all of the newly
30 synthesized polypeptide which is immunologically reactive with anti-CBHI, is produced under the direction of induced CBHI mRNA.

D.4 Preparation of a CBHI cDNA Probe

The enriched CBHI mRNA fraction from above provides a template for synthesizing radiolabeled, single-stranded CBHI cDNA for use as a probe in identifying genomic digest fragments of the selected
5 microbial organism which contain regions of the CBHI gene.

Methods for producing radiolabeled, single-stranded cDNA from mRNA are well known. The general method described in Payvar, F., et al., J. Biol. Chem. (1979) 254: 7636 was followed to produce cDNA from T. reesei strain L27 CBHI enriched mRNA. Briefly, the enriched
10 mRNA was pretreated with 10 mM methyl mercury hydroxide to denature the RNA, and then introduced into a reaction mixture containing radiolabelled nucleotides, oligo dT as a primer and 2 mM adenosine:$VOSO_4$ as an RNAase inhibitor. Following cDNA synthesis, the poly A RNA was destroyed by treatment with NaOH, and the cDNA was size
15 analyzed by gel electrophoresis to confirm the synthesis of full-length cDNA. A typical gel electrophoresis pattern of the cDNA fraction showed a prominent major band in the 1.6 to 1.8 kilobase size regions, corresponding approximately to the size range of the enriched mRNA fraction used to produce the cDNA.

20 A CBHI probe may also be produced, in accordance with known methods, by radiolabeling CBHI mRNA fragments. Alternatively, a synthetic polynucleotide constructed to have a codon sequence corresponding to the known amino acid sequence of a short segment of the mature CBHI protein may also provide a suitable CBHI probe for use
25 in the invention.

D.5 Preparation of Genomic DNA Clones Containing CBHI Gene Regions

Total genomic DNA isolated from the selected microbial source is treated with one or more selected endonucleases to generate genomic fragments which can be cloned and identified as containing
30 CBHI gene regions by their ability to hybridize the above cDNA CBHI probe. The clones identified as containing CBHI gene regions provide the gene material used in determining the nucleotide sequence of the genomic DNA, and clones containing selected regions of the gene may be

used in constructing an intron-free CBHI gene according to methods detailed below.

A complete library of T. reesei L27 DNA was constructed in lambda L47.1. The lambda phage vector was prepared by isolating phage DNA and purifying the cloning arms away from the stuffer fragment as follows: starting with whole phage DNA, cos sites were ligated together under "sticky end" conditions.   The efficiency of ligation was determined to be about 80-90% by agarose gel analysis. Ligated phage DNA was cut with BamHI and SalI. The BamHI digestion releases stuffer fragments from ligated arms, whereas SalI cuts the stuffer fragments twice more and insures greater ease of separation from the BamHI digested arms. The BamHI digested arms were purified from the stuffer fragments using a CsCl velocity sedimentation gradient.

T. reesei genomic DNA was partially digested with Sau3A and size fractionated on NaCl velocity gradients, and the fraction containing molecules in the 15-20 kilobase range selected. Fragments of this fraction were mixed with the BamHI arms of the vector in a 1:1 molar ratio, at a total DNA concentration of 200 ug/ml and ligated overnight at 4°C with 0.2 Weiss units of T4 DNA ligase in a volume of 20 microliters. One half microgram from the ligation was packaged into lambda heads in vitro, using frozen packaging components as described in Hohn, B., Methods in Enzymology (1979) 68:299-309, and the resulting phage were plated on E. coli strain MM 294. The procedure produced a library of about 1.7×10^4 phage, which gave a 0.99+ probability of any single copy gene being represented.

The T. reesei genomic library so produced was probed with radiolabeled CBHI cDNA prepared as above. Out of approximately 60,000 phage tested, 54 phage hybridized to the cDNA probe. To eliminate false positives, 23 clones which produced the strongest hybridization with the cDNA were further purified and probed with two different cDNA probes, one produced from mRNA from "induced" cultures, and the other from mRNA from "uninduced" cultures, lacking CBHI mRNA. 22 of the 23 selected genomic library clones hybridize the "induced" cDNA probe, but not the "uninduced" cDNA. Restriction

analysis of 15 of the 22 selected clones revealed several duplicates, reducing the CBHI gene library to six unique clones.

One of the six clones, designated lambda CBH-4, which was presumed to contain all of the CBHI genomic coding sequence was used to generate gene subfragment clones. Experimentally, lambda CBH-4 was digested to completion with HindIII and the digest fragments produced were ligated into HindIII-digested E. coli plasmid pBR322. Plasmids containing digested CBHI fragments in recombinant form were selected by ampicillin resistance (carried on the pBR322 vector), and screened for tetracycline sensitivity produced by insertion of heterologous DNA leading to inactivation of the tetracycline-resistance gene. Subclones containing either a 1.16 kilobase fragment, including a 5' region of the CBHI gene, or a 2.3 kilobase fragment, including a 3' portion of the gene, were selected. These two clones were designated pCBH157 and pCBH164, respectively, and are illustrated in Figures 3A and 3B, respectively.

pCBH157 contains the coding region for the first 278 amino acids of the mature CBHI protein from T. reesei strain L27. The heavy-lined segment of pCBH157 shown in Figure 3A represents the entire 1.16 kilobase DNA fragment, the interior line segment extending along a portion of the fragment represents the 5' CBHI gene region.

The pCBH164 subclone, containing the 2.3 kilobase HindIII fragment from lambda CBH-4, includes the region of the CBHI genomic gene which encodes the last 218 amino acids of the mature CBHI protein, from amino acids 279 to 496. The heavy-line segment in Figure 3B represents the 2.3 kilobase DNA fragment, and the interior line segment indicates the coding region of the CBHI genomic gene.

Figures 3A and 3B also indicate the positions of several restriction endonuclease sites within the CBHI gene region, as determined by conventional restriction endonuclease site analysis.

pCBH157 and pCBH164 were tested for their ability to "fish out" CBHI mRNA from a total poly A RNA preparation obtained from a cellulase-induced T. reesei L27 culture. Each clone was bound to a nitrocellulose filter, according to published procedures (Harpold, M.

H., et al., <u>Nucleic Acids Res.</u> (1978) <u>5</u>:2039) and total poly A RNA from an induced <u>T. reesei</u> L27 culture was added to the filters under conditions which allowed hybridization of complementary sequences. After stringent washes to remove unbound (nonspecific mRNA), the mRNA which specifically hybridized the filter-bound plasmid was eluted.

For each filter, bound and unbound RNA fractions were translated in the cell-free protein synthesizing system described above. Total proteins translated from the two samples were precipitated with a mixture of anti-CBHI antibodies and <u>Staphylococcus aureus</u> cells. The total translated proteins were analyzed on SDS-PAGE and by analysis of the immunoprecipitates as described above. Plasmids pCBH164 and pCBH157 both bound RNA capable of directing the synthesis of a protein which was immunologically reactive with anti-CBHI antibody.

D.6 <u>Determination of Genomic Sequences</u>

To obtain sequence information, subclones of the genomic CBHI gene, obtained in accordance with above-described methods, are further fragmented by digestion with one or more selected endonucleases to produce subcloned fragments suitable in length for nucleotide sequencing. By way of illustration, methods used in sequencing the CBHI genomic gene from <u>T. reesei</u> strain L27 are described below.

Subclone pCBH157, containing a 5' portion of the gene, was digested substantially to completion with both HindIII and EcoRI, producing two gene-region fragments of approximately 600 base pairs and 500 base pairs, as can be appreciated from Figure 3A. The two fragments were subcloned into M13 bacteriophage vectors M13mp8 and M13mp9, respectively.

The other HindIII fragment subclone, pCBH164, was digested variously with BamHI, HindIII, and HincII. The fragments produced were subcloned into the vector M13mp8.

The fragments of the CBHI genomic region subcloned into the vectors M13mp8 and M13mp9 were sequenced by the dideoxy chain

termination method described in Sanger, F., et al., <u>Nat. Acad. Sci.</u> <u>USA</u> (1977) <u>74</u>:5463 and Messing, J., et al., <u>Nucleic Acids Res.</u> (1981) <u>9</u>:309. Portions of the sequence were confirmed by the Maxam-Gilbert sequencing technique (Maxam, A. M., et al., <u>Proc. Nat. Acad. Sci. USA</u> (1977) <u>74</u>:560). The entire gene sequence, extending from the 5' HindIII site in the 1.16 kilobase <u>Hind</u>III fragment from <u>T. reesei</u> strain L27 to approximately 306 nucleotides past the stop codon at the 3' end of the CBHI gene in the 2.3 kilobase fragment, is shown in Table I below. The sequence data was internally consistent with respect to the various fragments sequenced.

The nucleotide sequence obtained was compared, in a computer-programmed matching operation, with the regions of known amino acid sequence of <u>T. reesei</u> CBHI from Fagerstam, L., et al. <u>supra</u> and Fagerstam, L. G. <u>supra</u>. The matching operation examined the nucleotide sequence in each of the three possible reading frames for codon correspondence with the given amino acid sequence. The matching operation produced nearly complete correspondence between coding regions of the CBHI gene and the regions of known amino acid sequence of CBHI from <u>T. reesei</u>, QM 9414. This matching allowed the identification of the correct reading frame, the start and stop codons, and the two introns in the gene, as will be discussed. Table I also shows the amino acids corresponding to the leader sequence encoding portions of the gene. (Leader sequence amino acids are underlined.) Recognition sequences of selected restriction endonucleases referred to in the text are indicated by underlining in the table. Overlined portions of the untranslated introns indicate recognizable sequences analogous to those of introns reported in other genes.

TABLE I

AAGCTTAGCC AAGAACAATA GCCGATAAAG ATAGCCTCAT TAAACGGAAT
1 HindIII 20 30 40 50

GAGCTAGTAG GCAAAGTCAG CGAATGTGTA TATATAAAGG TTCGAGGTCC
60 70 80 90 100

5 GTGCCTCCCT CATGCTCTCC CCATCTACTC ATCAACTCAG ATCCTCCAGG
110 120 130 140 150

AGACTTGTAC ACCATCTTTT GAGGCACAGA AACCCAATAG TCAACCGCGG
160 170 180 HincII SacII

		MET	TYR	ARG	LYS	LEU	ALA	VAL	ILE	SER	ALA	PHE
10	ACTGGCATC	ATG	TAT	CGG	AAG	TTG	GCC	GTC	ATC	TCG	GCC	TTC
	209				224						BglI	

LEU	ALA	THR	ALA	ARG	ALA	GLN	SER	ALA	CYS	THR	LEU	GLN	SER
TTG	GCC	ACA	GCT	CGT	GCT	CAG	TCG	GCC	TGC	ACT	CTC	CAA	TCG
		254						269					284

15	GLU	THR	HIS	PRO	PRO	LEU	THR	TRP	GLN	LYS	CYS	SER	SER	GLY
	GAG	ACT	CAC	CCG	CCT	CTG	ACA	TGG	CAG	AAA	TGC	TCG	TCT	GGT
					299					314				

	GLY	THR	CYS	THR	GLN	GLN	THR	GLY	SER	VAL	VAL	ILE	ASP	ALA
	GGC	ACT	TGC	ACT	CAA	CAG	ACA	GGC	TCC	GTG	GTC	ATC	GAC	GCC
20	329					344					359			

	ASN	TRP	ARG	TRP	THR	HIS	ALA	THR	ASN	SER	SER	THR	ASN	CYS
	AAC	TGG	CGC	TGG	ACT	CAC	GCT	ACG	AAC	AGC	AGC	ACG	AAC	TGC
		HaeII						389				404		

```
    TYR ASP GLY ASN THR TRP SER SER THR LEU CYS PRO ASP ASN
    TAC GAT GGC AAC ACT TGG AGC TCG ACC CTA TGT CCT GAC AAC
            419                     434                 449


    GLU THR CYS ALA LYS ASN CYS CYS LEU ASP GLY ALA ALA TYR
 5 · GAG ACC TGC GCG AAG AAC TGC TGT CTG GAC GGT GCC GCC TAC
            464                     479                 494


    ALA SER THR TYR GLY VAL THR THR SER GLY ASN SER LEU SER
    GCG TCC ACG TAC GGA GTT ACC ACG AGC GGT AAC AGC CTC TCC
                    509                 524


10  ILE GLY PHE VAL THR GLN SER ALA GLN LYS ASN VAL GLY ALA
    ATT GGC TTT GTC ACC CAG TCT GCG CAG AAG AAC GTT GGC GCT
    539                     554     MstI         569


    ARG LEU TYR LEU MET ALA SER ASP THR THR TYR GLN GLU PHE
    CGC CTT TAC CTT ATG GCG AGC GAC ACG ACC TAC CAG GAA TTC
15       ·584             599                 614     EcoRI


    THR LEU LEU GLY ASN GLU PHE SER PHE ASP VAL ASP VAL SER
    ACC CTG CTT GGC AAC GAG TTC TCT TTC GAT GTT GAT GTG TCG
            629                 644                 659


    GLN LEU PRO   _____
20 · CAG CTG CCG  TAAGTGACTT  ACCATGAACC  CCTGACGTAT  CTTCTTGTGG
    PvuII                 681         691         701         711


              _____        _____  CYS GLY LEU ASN GLY ALA
    GCTCCCAGCT GACTGGCCAA  TTTAAGG TGC GGC TTG AAC GGA GCT
         PvuII          731                             753


25  LEU TYR PHE VAL SER MET ASP ALA ASP GLY GLY VAL SER LYS
    CTC TAC TTC GTG TCC ATG GAC GCG GAT GGT GGC GTG AGC AAG
            768     NcoI             783                 798
```

	TYR	PRO	THR	ASN	THR	ALA	GLY	ALA	LYS	TYR	GLY	THR	GLY	TYR
	TAT	CCC	ACC	AAC	ACC	GCT	GGC	GCC	AAG	TAC	GGC	ACG	GGG	TAC
					813		HaeII				829			

	CYS	ASP	SER	GLN	CYS	PRO	ARG	ASP	LEU	LYS	PHE	ILE	ASN	GLY
5	TGT	GAC	AGC	CAG	TGT	CCC	CGC	GAT	CTG	AAG	TTC	ATC	AAT	GGC
			843					858			873			

	GLN	ALA	ASN	VAL	GLU	GLY	TRP	GLU	PRO	SER	SER	ASN	ASN	ALA
	CAG	GCC	AAC	GTT	GAG	GGC	TGG	GAG	CCG	TCA	TCC	AAC	AAC	GCA
		888					903					918		

	ASN	THR	GLY	ILE	GLY	GLY	HIS	GLY	SER	CYS	CYS	SER	GLU	MET
10	AAC	ACG	GGC	ATT	GGA	GGA	CAC	GGA	AGC	TGC	TGC	TCT	GAG	ATG
			933				948					963		

	ASP	ILE	TRP	GLU	ALA	ASN	SER	ILE	SER	GLU	ALA	LEU	THR	PRO
15	GAT	ATC	TGG	GAG	GCC	AAC	TCC	ATC	TCC	GAG	GCT	CTT	ACC	CCC
				978					993					1008

	HIS	PRO	CYS	THR	THR	VAL	GLY	GLN	GLU	ILE	CYS	GLU	GLY	ASP
	CAC	CCT	TGC	ACG	ACT	GTC	GGC	CAG	GAG	ATC	TGC	GAG	GGT	GAT
					1023				BglII					

	GLY	CYS	GLY	GLY	THR	TYR	SER	ASP	ASN	ARG	TYR	GLY	GLY	THR
20	GGG	TGC	GGC	GGA	ACT	TAC	TCC	GAT	AAC	AGA	TAT	GGC	GGC	ACT
						1068					1083			

	CYS	ASP	PRO	ASP	GLY	CYS	ASP	TRP	ASN	PRO	TYR	ARG	LEU	GLY
	TGC	GAT	CCC	GAT	GGC	TGC	GAC	TGG	AAC	CCA	TAC	CGC	CTG	GGC
		1098					1113					1128		

	ASN	THR	SER	PHE	TYR	GLY	PRO	GLY	SER	SER	PHE	THR	LEU	ASP
25	AAC	ACC	AGC	TTC	TAC	GGC	CCT	GGC	TCA	AGC	TTT	ACC	CTC	GAT
			1143					1158	HindIII				1173	

```
    THR  THR  LYS  LYS  LEU  THR  VAL  VAL  THR  GLN  PHE  GLU  THR  SER
    ACC  ACC  AAG  AAA  TTG  ACC  GTT  GTC  ACC  CAG  TTC  GAG  ACG  TCG
                   1188                     1203                     1218


    GLY  ALA  ILE  ASN  ARG  TYR  TYR  VAL  GLN  ASN  GLY  VAL  THR  PHE
 5· GGT  GCC  ATC  AAC  CGA  TAC  TAT  GTC  CAG  AAT  GGC  GTC  ACT  TTC
                        1233                     1248


    GLN  GLN  PRO  ASN  ALA  GLU  LEU  GLY  SER  TYR  SER  GLY  ASN  GLU
    CAG  CAG  CCC  AAC  GCC  GAG  CTT  GGT  AGT  TAC  TCT  GGC  AAC  GAG
    1263                1278                     1293


10· LEU  ASN  ASP  ASP  TYR  CYS  THR  ALA  GLU  GLU  ALA  GLU  PHE  GLY
    CTC  AAC  GAT  GAT  TAC  TGC  ACA  GCT  GAG  GAG  GCA  GAA  TTC  GGC
                             ‾‾‾‾‾‾‾‾                   ‾‾‾‾‾‾‾‾
                                  PvuII                      EcoRI


    GLY  SER  SER  PHE  SER  ASP  LYS  GLY  GLY  LEU  THR  GLN  PHE  LYS
    GGA  TCC  TCT  TTC  TCA  GAC  AAG  GGC  GGC  CTG  ACT  CAG  TTC  AAG
    ‾‾‾‾‾‾‾‾
15· BamHI     1353                     1368                     1383


    LYS  ALA  THR  SER  GLY  GLY  MET  VAL  LEU  VAL  MET  SER  LEU  TRP
    AAG  GCT  ACC  TCT  GGC  GGC  ATG  GTT  CTG  GTC  ATG  AGT  CTG  TGG
                   1398                     1413                     1428


    ASP  ASP  ‾‾‾‾‾‾‾
20  GAT  GAT  GTGAGTTTGA  TGGACAAACA  TGCGCGTTGA  CAAAGAGTCA
                  1444        1454        HincII       1474


              ‾‾‾‾‾‾‾        ‾‾‾‾‾‾‾       TYR  TYR  ALA  ASN  MET  LEU  TRP
    AGCAGCTGAC  TGAGATGTTA  CAG  TAC  TAC  GCC  AAC  ATG  CTG  TGG
       PvuII           1494                          1512


    LEU  ASP  SER  THR  TYR  PRO  THR  ASN  GLU  THR  SER  SER  THR  PRO
25· CTG  GAC  TCC  ACC  TAC  CCG  ACA  AAC  GAG  ACC  TCC  TCC  ACA  CCC
                   1527                1542                     1557
```

GLY	ALA	VAL	ARG	GLY	SER	CYS	SER	THR	SER	SER	GLY	VAL	PRO
GGT	GCC	GTG	CGC	GGA	AGC	TGC	TCC	ACC	AGC	TCC	GGT	GTC	CCT
			1572					1587					1602

ALA	GLN	VAL	GLU	SER	GLN	SER	PRO	ASN	ALA	LYS	VAL	THR	PHE
GCT	CAG	GTC	GAA	TCT	CAG	TCT	CCC	AAC	GCC	AAG	GTC	ACC	TTC
5				1617				1632		BstEII			

SER	ASN	ILE	LYS	PHE	GLY	PRO	ILE	GLY	SER	THR	GLY	ASN	PRO
TCC	AAC	ATC	AAG	TTC	GGA	CCC	ATT	GGC	AGC	ACC	GGC	AAC	CCT
1647					1662				1677				

SER	GLY	GLY	ASN	PRO	PRO	GLY	GLY	ASN	ARG	GLY	THR	THR	THR
AGC	GGC	GGC	AAC	CCT	CCC	GGC	GGA	AAC	CGT	GGC	ACC	ACC	ACC
10		1692				1707				1722			

THR	ARG	ARG	PRO	ALA	THR	THR	THR	GLY	SER	SER	PRO	GLY	PRO
ACC	CGC	CGC	CCA	GCC	ACT	ACC	ACT	GGA	AGC	TCT	CCC	GGA	CCT
15		1737				1752				1767			

THR	GLN	SER	HIS	TYR	GLY	GLN	CYS	GLY	GLY	ILE	GLY	TYR	SER
ACC	CAG	TCT	CAC	TAC	GGC	CAG	TGC	GGC	GGT	ATT	GGC	TAC	AGC
			1782				1797					1812	

GLY	PRO	THR	VAL	CYS	ALA	SER	GLY	THR	THR	CYS	GLN	VAL	LEU
GGC	CCC	ACG	GTC	TGC	GCC	AGC	GGC	ACA	ACT	TGC	CAG	GTC	CTG
20				1827				1842					

ASN	PRO	TYR	TYR	SER	GLN	CYS	LEU
AAC	CCT	TAC	TAC	TCT	CAG	TGC	CTG	TAAAGCTCCG	TGCGAAAGCC
1857				1872				1888	1898

TGACGCACCG   GTAGATTCTT   GGTGAGCCCG   TATCATGACG   GCGGCGGGAG

25   1908   1918   1928   1938   1948

```
CTACATGGCC  CCGGGTGATT  TATTTTTTTT  GTATCTACTT  CTGACCCTTT
        SmaI    1968            1978        1988        1998


TCAAATATAC  GGTCAACTCA  TCTTTCACTG  GAGATGCGGC  CTGCTTGGTA
    2008    HincII          2028        2038        2048


TTGCGATGTT  GTCAGCTTGG  CAAATTGTGG  CTTTCGAAAA  CACAAAACGA
    2058        2068        2078        2088        2098


TTCCTTAGTA  GCCATGCATT  TTAAGATAAC  GGAATAGAAG  AAAGAGGAAA
    2108        2118        2123        2138        2148


TTAAAAAAAA  AAAAAAAACA  AACATCCCGT  TCATAACCCG  TAGAATCGCC
    2158        2168        2178        2188        2198


GCTCTTCGTG  TATCCCAGTA  CCA
    2208        2218
```

.5

10

Viewing the table, it is seen that ATG start codon is located at nucleotides 210-212, numbered from the 5'-end of the inserted sequence in pCBH157. The codon at nucleotides 261-263 corresponds to the N-terminus glutamine (pyroglutamine) amino acid in the mature CBHI protein. The codons between nucleotides 210 and 260 presumably code a 17-amino acid hydrophobic peptide leader sequence whose amino acid sequence, as determined by corresponding codon sequence, is shown underlined in the table. The codons between nucleotides 261 and 671 code for the first 137 amino acids of the mature CBHI enzyme.

Between the coding regions coding for amino acids 137 and 138 in the mature CBHI protein, is a 67 base pair region which is identified herein as the first, or upstream intron in the CBHI gene. This intron contains a 7 base-pair AGCTGAC sequence (highlighted by overlining in Table I), which is similar to a 7-mer in the intron sequence of the actin gene from yeast (Langford, C. J., et al., Cell (1983) 33:519) and in three of the introns in the glucoamylase gene

from Aspergillus awamori and which is identical to an intron sequence in the glutamate dehydrogenase gene from Neurospora crassa (Kinnaird, J., et al., Gene (1982) 20: 387) and one of the introns in the A. awamori glucoamylase gene. This sequence apparently provides a signal which is necessary for splicing the intron out of the mRNA transcript. In addition, the intron contains donor and acceptor consensus sequences TAAGTG and TTTAAGG, respectively (also highlighted by overlining in Table I), which are identical to donor and acceptor site consensus sequences which have been cataloged by Mount, S. M., Nucleic Acids Res. (1982) 10:459.

Amino acid 138 (cysteine) through amino acid 369 (aspartate) are encoded by an uninterrupted protein-encoding region extending between nucleotides 739-1434. The gene region between the codons coding for amino acids 369 and 370 contain 63 non-coding nucleotides which make up a second, or downstream intron in the genomic CBHI gene. The second intron also contains the 7-base pair sequence AGCTGAC characteristic of yeast introns, and donor and acceptor consensus sequences GTGAGT, and TTACAG, respectively, which are identical to known donor and acceptor site sequences at intron-exon junctions in yeast genes (Mount, S. M., supra).

The second intron is followed by an uninterrupted 127-codon region encoding for amino acids 370-496 of the mature CBHI protein, whose C-terminal amino acid is leucine (encoded by nucleotides 1876-1878). Nucleotides 1879-2221, which are downstream of stop codon at the 3' coding region of the gene, contain SmaI and HincII sites as shown.

The sequence of nucleotides in the protein-encoding portions of the gene in Table I forms a codon sequence which encodes the sequence of amino acids shown. Many of the codons can be varied in nucleotide sequence, generally at the third nucleotide position, without changing the amino acid specified by the codon. A variety of methods for producing single nucleotide, or point mutations in genes are known to those in the art. The term "codon sequence", as defined herein, includes all nucleotide sequences which encode a particular

amino acid sequence, the nucleotide sequence shown in Table I being one of the many codon sequences which encodes the amiono acid sequence there shown.

In addition to nucleotide substitutions which preserve the
5 gene's codon sequence, the present invention contemplates nucleotide variations in the gene which lead to "neutral" or non-critical amino acid changes in the encoded CBHI enzyme, as discussed above. It is known, particularly from studies on evolutionary changes in protein amino acid sequences, that neutral amino acid changes in a protein can
10 occur without significantly changing the protein's functional characteristics. A variety of techniques are available for producing minor codon changes experimentally, including mutagenizing a gene to produce point or deletion mutations; selectively cleaving, followed by deleting or adding selected nucleotides, and then religating the gene;
15 and by applying conventional oligonucleotide mutagenesis techniques. A CBHI gene so mutated may encode an enzymatically active CBHI enzyme, and often one whose specific activity is substantially unchanged with respect to the unmutated CBHI enzyme. The present invention includes genes whose codon sequences encode such neutral amino acid changes in
20 the CBHI enzyme, as well as those whose codon sequences have strict one-to-one codon correspondences with the amino acid sequence of an unmutated mature CBHI. As defined herein, a gene whose codon sequence has a "substantially" one-to-one correspondence with mature CBHI includes both those having a strict correspondence and as those having
25 neutral or minor codon variations which encode a polypeptide with CBHI (exo-cellobiohydrolase) enzyme activity (see Shoemaker, S. P., et al., Biochem. Biophys. Acta (1978) 523:147).

Figure 4 shows a map of the genomic CBHI gene from T. reesei strain L27, constructed from the above sequence data. The map shows
30 the protein-encoding region of the gene in relation to several restriction endonuclease sites, and in particular, in relation to the HindIII sites defining the above-described 1.16 and 2.3 kilobase HindIII gene fragments. As seen from the map, the upstream intron in the gene is located in the central region of the 1.16 kilobase HindIII
35 fragment, and the downstream intron, in the central protein-encoding

portion of the 2.3 kilobase fragment.

In the example just discussed, the existence and location of the upstream and downstream introns in the <u>T. reesei</u> strain L27 CBHI gene were determined by comparing the codon sequence of the sequenced genomic gene with the amino acid sequence of the encoded protein, a method which can be used to define the introns with high precision. Alternatively, the lengths and positions of the gene introns can be determined by comparing the nucleotide sequence of the genomic CBHI gene with that of a full-length CBHI cDNA prepared and sequenced in accordance with the general procedures described herein. It may also be possible to identify the existence and location of genomic gene introns from the nucleotide sequence alone. As discussed above, both introns in the <u>T. reesei</u> strain L27 CBHI gene contain a 7 base-pair sequence characteristic of yeast introns, thus providing a means for making a preliminary identification of the introns. Similarly, the donor and acceptor end regions of the intron may be identified by characteristic consensus sequences which match one of the several consensus sequences which have been cataloged. While the invention has been illustrated with reference to the amino acid-codon matching method of intron identification, the other two just-mentioned methods for determining the existence and location of gene introns from the nucleotide sequence are also included in the method of the invention.

D.7. <u>Elimination of Intron Sequences From the Gene</u>

This section describes a method for constructing an intron-free CBHI gene. The method includes, first, providing in a suitable cloning vector, a 5' genomic CBHI gene fragment which extends from the 5' gene end to a selected restriction endonuclease site downstream of the 5'-end intron in the genomic CBHI gene. Generally, the methods described above in subparagraph D.5 for producing genomic DNA digest clones containing CBHI gene regions are applicable for producing such a 5' genomic gene fragment.

In particular, one advantageous procedure illustrated herein involves a first step of treating the genomic DNA with a selected restriction endonuclease under conditions which produce one or more

fragments containing the entire CBHI gene. This complete-gene fragment is then further digested to generate two or more gene subfragments, including, for example the 1.16 kilobase 5' CBHI gene fragment from T. reesei, strain L27 described above.

5   The second step includes providing, in recombinant form, one or more CBHI cDNA coding sequence fragments which can be spliced together or substituted for corresponding intron containing regions of genomic gene fragments, to produce uninterrupted protein-encoding gene regions. The cDNA gene fragments are constructed by conventional

10   methods, preferably by copying an enriched CBHI mRNA to form a single-stranded cDNA, and forming a double-stranded cDNA from the single strand.

The general prodecure for providing enriched CBHI mRNA follows that described in subparagraph D.3 above. The enriched mRNA,

15   when reacted with oligo dT in the presence of reverse transcriptase, serves as a template for synthesizing a cDNA strand, also according to above-described methods. The mRNA in the mRNA/cDNA hybrid is hydrolyzed by treatment with NaOH, leaving a single-stranded cDNA which can serve to prime and as a template for the synthesis of a

20   second cDNA strand by E. coli DNA polymerase 1, according to methods described in Maniatis, et al., Molecular Cloning - a Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York (1982) pp. 235-238, to give a "hairpin" of double stranded DNA. Following synthesis of the second cDNA strand, the hairpin loop may be

25   removed by treatment with nuclease S1, and various duplex cDNA's selected.

Double-stranded cDNA obtained by copying T. reesei strain L27 CBHI mRNA was formed according to above-described methods and cloned into a pUC8 vector according to standard procedures. The

30   clones were identified by their ability to hybridize one of the above-described pCBH157 or pCBH164 clones containing regions of CBHI genomic DNA. One of the cDNA fragment clones selected, designated p805, is shown in Figure 6. The cDNA fragment in p805, which is indicated by heavy line in the figure, extends from the 5'-end of the gene

downstream past an NcoI site which, in the genomic gene, is located on the downstream side of the first or upstream intron, as indicated.

Two other cDNA clones, identified as p417 and p247, are seen in Figure 7. As shown, the cDNA fragment in p247 extends from the 3'-
5 end of the cDNA upstream past a BstEII site. The cDNA fragment in p417 overlaps fragment 247 at the just-mentioned BstEII site, and extends upstream therefrom past the internal HindIII site in the CBHI gene, as shown. The lengths and relative positions of the three just-described cDNA fragments were determined by conventional restriction
LO endonuclease mapping.

To construct the intron-free CBHI gene, a portion of the genomic DNA containing the the upstream intron was excised from 5' genomic fragment and a corresponding cDNA region inserted into the excised intron-containing region. This intron-free 5' gene fragment
L5 was then joined, at a selected restriction endonuclease site, to a 3'-end cDNA fragment which extends from such site to the 3'-end of the gene. In the illustration below, the intron-free 5' gene fragment was cloned into a suitable expression vector adjacent control sequences which allow expression of the fragment when the vector is introduced
20 into a suitable host. The remaining 3' cDNA portion of the gene was then introduced into the expression vector to produce a full-length, intron-free CBHI gene positioned and oriented in the vector for gene expression.

D.7.a. Construction of pCBH5--Elimination of the Upstream Intron

25 Figure 6 illustrates plasmid pCBH157, containing the 1.16 kilobase HindIII fragment from T. reesei strain L27 genomic DNA which is also shown in Figure 3A. From the gene nucleotide sequence data presented above, it is known that the 1.16 kilobase fragment includes a unique SacII recognition site 11 base pairs 5' to the start
30 of translation at the first ATG initiation signal in the gene, and a unique NcoI cleavage site 32 base pairs 3' to the upstream intron in the gene fragment, as indicated in the figure.

Plasmid p805, containing an approximately 800 base pair 5'-
end cDNA fragment, also contains the just-mentioned unique SacII and
NcoI sites, as seen in Figure 6. Plasmids pCBH157 and p805 were each
treated with SacII and NcoI, releasing from plasmid pCBH157 a 573 base

5   pair fragment containing the upstream intron, and from plasmid 805, a
corresponding intron-free fragment 506 base pairs in length. The
intron-free SacII/NcoI fragment from p805 was ligated into the
digested pCBH157 plasmid to produce a new plasmid, designated pCBH5,
containing a chimeric genomic/cDNA gene fragment consisting of the

10   1.16 kilobase HindIII fragment from which the upstream intron has been
eliminated.

D.7.b. Construction of pCBH3--Elimination of the Downstream Intron

Construction of a 3' intron-free gene fragment will now be
considered. With reference to Figure 7, plasmid p247 contains a SmaI

15   site 77 base pairs distal to the TAA stop codon of the gene; a unique
BstEII site approximately 138 nucleotides 3' to the downstream intron
in the genomic sequence and a unique SalI site in the polylinker
region of the plasmid, 5' with respect to the cDNA coding sequence.

Plasmid p417, also illustrated in Figure 7, contains the

20   same unique BstEII site as in p247 but the CBH coding sequence and
extends further in the 5' direction past the HindIII recognition site
within the CBHI coding sequence, and includes a unique SalI site about
600 base pairs upstream of the BstEII site, as shown.

To construct a composite 3' cDNA gene clone extending from

25   the downstream end of the gene's stop codon upstream past the internal
HindIII site, plasmids p247 and p417 were each digested to completion
with BstEII and then SalI. The smaller 600 base pair BstEII/SalI DNA
fragment from p417 was purified and ligated with the BstEII/SalI p247
vector fragment. Plasmid pCBH3 shown in the figure was identified as

30   the desired recombinant having a cDNA fragment extending from the 3'-
end of the cDNA to the internal HindIII site in the gene.

D.8 Construction of Expression Vectors for Bacterial Hosts

The CBHI gene fragments from pCBH5 and pCBH3 were cloned into a selected E. coli/S. cerevisiae shuttle vector to a position and orientation adapted to allow CBHI gene expression under control of the trp promoter. The host vector is a 11.12 kilobase shuttle vector designated pDG151 deposited at ATCC on May 11, 1984 and designated accession number 39686, detailed in Figure 5. The plasmid includes a modified aminoglycoside phosphotransferase APH-I gene which encodes an enzyme which confers antibiotic resistance against kanamycin, neomycin, G418, and other antibiotics.

The sequences of this plasmid are as follows:

1. Coordinates 0-1.54 are the 1.54 kb HindIII/EcoRI 3' untranslated terminator sequences of the EN01 gene derived from peno46 (Holland, M. J., et al., J. Biol. Chem. (1981) 256: 1385).

2. Coordinates 1.54-2.75 contain the modified APH-I gene, which is described in detail in The Molecular Biology of Yeast, Cold Spring Harbor Meeting, August 16-21, 1983.

3. Coordinates 2.75-2.85 contain a sequence comprising duplicated lac operator sequences flanked by inverted polylinker repeats. The sequence immediately preceding the ATG start codon of the modified APH-I (at 2.75) is:

```
5' G AAT TCC CGG GGA TCC GGC GAT AAG CTT ATG
       EcoRI        BamHI          HindIII

           --------
            XmaI
```

4. Coordinates 2.85-2.95 are a 107 bp 5'- EcoRI(repaired)/BamHI-3' fragment containing the trp promoter- operator.

5. Coordinates 2.95-3.04 contain a 90 bp pBR322 segment between the SphI(repair) and SalI sites.

6. The LEU2 gene from yeast occupies coordinates 3.04- 5.25. It was obtained as an XhoI/SalI digest fragment from YEp13 (Broach, J., et al., Gene (1979) 8: 121).

7. The 2 micron plasmid replicon in coordinates 5.25-8.97 is obtained by digestion of pDB248 (Beach, D., et al., _Nature_ (1981) __290__: 140) with EcoRI (repair) and SalI, and isolation of the 2.7 kb DNA fragment containing the replicon. The existence of the

5 appropriate SalI site was not deducible from the disclosure of Beach. However, pDB248 was shown to contain a SalI site about 50 bp downstream from the indicated LEU2 region/2μ PstI tailing site as set forth in the Beach reference.

8. Coordinates 8.97-11.12 contain a TthlllI(repair)/EcoRI

10 digest fragment of pBR322 which supplies Amp^R and an _E. coli_ origin of replication.

Figure 8 illustrates the method by which the intron-free 5'-end gene fragment from pCBH5 was inserted in pDG151. Plasmid pCBH5 was digested to completion with SacII, and the 3' protruding two-base

15 sticky end was repaired with _E. coli_ DNA polymerase I (Klenow) in the presence of dCTP. Following inactivation of the PolI, the DNA substrate was digested to completion with HindIII and the 895 base pair SacII(repair)/HindIII fragment was purified. With reference to the expanded gene region shown in Figure 5, plasmid pDG151 was

20 digested to completion with XmaI, liberating a 62 base pair DNA fragment containing the duplicated synthetic _lac_ operator fragment between the two adjacent XmaI sites. The 5' protruding four-base cohesive ends were repaired with PolI, in the presence of dCTP and dGTP. Following inactivation of the PolI, the flush ended DNA was

25 digested to completion with HindIII. The repaired XmaI site in the 11.12 kilobase vector DNA fragment is 18 nucleotides 3' of the _E. coli_ trp promoter and leader peptide ribosome binding site, and the cohesive HindIII site immediately precedes the initiating codon of the modified APH-I gene.

30 The pDG151 DNA fragments (1.5 micrograms) and the purified pCBH5 DNA fragment (0.24 micrograms), were ligated under sticky-end conditions at a total DNA concentration of about 60 micrograms per ml. The ligated linear DNA fragments were diluted to 20 micrograms per ml and ligation was continued under blunt-end conditions to favor

35 intramolecular circle formation. _E. coli_ K12 strain GM119 was

transformed to ampicillin resistance with 150 ng of the ligated DNA and non-constitutive Lac+ colonies were screened for the presence of the desired 11.94 kilobase plasmid, ptrpCBH5. Plasmid ptrpCBH5, shown in Figure 8, contained the desired insert with a new unique XmaI

5   recognition site at the repaired pDG151 XmaI site. The plasmid construction shown in Figure 8 was confirmed by restriction enzyme mapping. Plasmid ptrpCBH5 in <u>E. coli</u> K12/GM119 has been deposited in the Cetus Master Culture Collection (CMCC) and has CMCC deposit #1842. The plasmid has been sent to the Northern Regional Research

10 Laboratory (NRRL), Peoria, IL for deposit, and has NRRL #B15574.

D.8.a.   <u>Construction of ptrpCBH8</u>

The procedure for subcloning the 3' cDNA gene fragment from pCBH3 into ptrpCHB5 is illustrated in Figure 9. Plasmid pCBH3 was digested to completion with HindIII, releasing an approximately 800

15 base pair 3' CBHI coding fragment which spans from the internal HindIII site of the CBHI gene past the 3'-end of the gene adjacent the SmaI site, to include a sequence from pUC8 adjacent the 3' HindIII site. Plasmid ptrpCBH5 was also digested with HindIII to linearize the vector at the HindIII site immediately preceding the initiation

20 codon at the APH-I gene. The purified pCBH3/HindIII fragment (0.2 micrograms) was ligated under sticky end conditions with the ptrpCBH5/HindIII fragments (1.6 micrograms). <u>E. coli</u> K12 strain MM294 was transformed to ampicillin resistance, and colonies were screened for the presence of the desired plasmid. The construction of the new

25 plasmid, designated ptrpCBH8, is shown in Figure 9. Restriction mapping with HindIII, BamHI and EcoRI confirmed the construction shown. Plasmid ptrpCBH8 contains two XmaI/SmaI sites flanking the CBHI coding sequence. Digestion of ptrpCBH8 with either enzyme yields a 1.63 kilobase DNA CBHI gene fragment containing the entire CBHI

30 coding region about 12 base pairs at the 5'-end of the coding region and about 80 base pairs at the 3'-end of the coding region. Plasmid ptrpCBH8 in <u>E. coli</u> K12/MM294 is identified as CMCC #1841, and has been deposited with NRRL under the No. NRRL #B15573.

D.8.b. Construction of ptrpCBH81

The procedure for inserting the 3' CBHI cDNA fragment from pCBH3 into ptrpCBH5 is as follows: pCBH3 was treated with SmaI and HindIII to release a 734 base pair fragment containing the entire
5   coding region of the CBHI gene downstream of the internal HindIII site. The shuttle vector ptrpCBH5 (prepared in GM119, an E. coli K12 dam⁻ host) was treated successively with endonucleases StuI and HindIII to release a HindIII/StuI vector fragment. The purified C3H3 SmaI/HindIII fragment from pCH3 was ligated under sticky-end then
10 blunt end conditions with the ptrpCBH5 StuI/HindIII fragments. E. coli K12/MM294 was transformed to ampicillin resistance, and colonies were screened for the presence of the desired plasmid. The construction of the new plasmid, designated ptrpCBH81 is shown in Figure 10. Plasmid ptrpCBH81 in E. coli K12/MM294 is identified as
15 CMCC #1843, and has been sent to NRRL for deposit, where it has NRRL ≠B15575.

As indicated in Figure 10, the SmaI/StuI fusion does not regenerate an SmaI site in the new vector. The ptrpCBH81 plasmid thus has a unique SmaI/XmaI site 12 base pairs preceding the initiation
20 (methionine) codon of the CBHI intron-free gene, and 16 base pairs following the trp leader ribosome binding site.

D.8.c Construction of ptrpCBH82

The plasmids ptrpCBH8 and ptrpCBH81 described in paragraphs D.8.a and D.8.b above were further modified to decrease the spacing
25 between the trp ribosome binding site (RBS) and the ATG start codon of the desired CBHI encoding sequence. As detailed below, 23 nucleotides of the intervening sequence between the RBS and ATG initiating codon in a ptrpCBH8 fragment containing this sequence were replaced by a seven nucleotide segment, and the resulting modified fragment
30 substituted for the corresponding portion of ptrpCBH81. The altered sequence further contains overlapping ClaI and AhaIII recognition sites which replace BamHI and XmaI sites previously contained in this region. The replacement procedure is described below with reference

to Figure 11.

D.8.c.1. Modification of the pre-ATG sequence

A SalI/PstI DNA fragment was excised from ptrpCBH8 by double digestion with these enzymes. This fragment extends from the SalI site on the 5' side of the pDG151 trp promoter (Figure 5) to the PstI site adjacent the 3' side of the CBH gene (Figure 7). The fragment was isolated by agarose gel electrophoresis and recovered by electroelution.

Bacteriophage M13mp10w DNA was treated with SalI, followed by digestion with PstI to release a SalI/PstI phage fragment. Bacteriophage M13mp10w is a conventional phage of the type used for site mutagenesis; comparable types of phage are available from New England Biolabs (Beverly, MA). The SalI/PstI fragment was isolated by agarose gel electrophoresis and recovered by electrolution.

The ptrpCBH8 and M13mp10w SalI/PstI fragments prepared above were ligated under standard conditions and transfected into frozen competent E. coli K12 strain DG98. The cells were plated on media containing 5×10^{-4} M isopropyl thiogalactoside (IPTG) obtained from Sigma Chem. (St. Louis, MO.) and 40 μg/ml X-gal . Non α-complementing white plaques (16/405, 4%) were picked into fresh media. Mini cultures were screened for recombinant single strand phage DNA of the expected (9 kb) size. The structure of the desired recombinant phage, designated A6, was confirmed using restriction analysis.

A chemically synthesized, purified, oligodeoxyribonucleotide having the sequence:

5'-CAACCTCCGATACATTTTAAATCGATACCCTTTTTAC-3'

was radiolabelled according to a modification of the technique of Maxam and Gilbert (Maxam, A., et al., Methods in Enzymology (1980) 68: 521, Academic Press). Briefly, 87 picomoles (2900 Ci/mmol) of ^{32}p-ATP obtained from New England Nuclear, were dried and combined with 50 picomoles of the oligomer and 12U of T4 polynucleotide kinase in a total volume of 30 μl containing 40 mM Tris-Cl, pH 7.6, 10 mM MgCl$_2$, 5 mM DTT, 0.1 mM spermidine, and 0.1 mM EDTA. The mixture was incubated

for one hour at 37°C, another 12U of kinase was added and the reaction was continued for an additional hour. The crude mixture was purified by PAGE using a 1 mm, 16% acrylamide gel, electrophoresed at 500V for 30 min. with cooling. The band with the slowest migration was
5 excised, crushed, and eluted twice with 0.5 ml hybridization buffer after heating at 100°C for 10 min. The specific activity of the solution was 5.3×10^6 dpm/pm (75% incorporation).

Recombinant M13mp10w bacteriophage A6 was prepared in _E. coli_ K12 strain DG98 and the single strand phage DNA purified. One
10 pmole of single strand phage DNA and 10 pmoles of the above synthetic nucleotide primer (not kinased) were annealed by heating for 1 min. at 67°C, and then 30 min. at 37°C in 15 µl 20 mM Tris-Cl, pH 8, 20 mM $MgCl_2$, 100 mM NaCl, 20 mM 2-mercaptoethanol. The annealed DNA was incubated with DNA polymerase I (Klenow) and 500 µM dNTPs for 30 min.,
15 0°C and then brought to 37°C. Aliquots (0.05 or 0.25 pmole) were removed after 5 min., 20 min., and 45 min., transformed into _E. coli_ K12 strain MM294 and plated with _E. coli_ K12 strain DG98.

The plates were chilled at 4°C and plaques lifted with PalI membranes obtained from Biodyne (1-2 min. in the first filter, more
20 than 10 min. for the second filter. The filters were denatured in 2.5 M NaCl, 0.5 M NaOH (5 min.). The denaturing medium was neutralized with 3 M sodium acetate to pH 5.5, the filters baked at 80°C _in vacuo_ for one hour, and then prehybridized in 6X SSC, 5X Denhardt's solution, 0.1% SDS, 50 µg/ml yeast t-RNA at 54.4°C, two hours. The
25 filters were then probed with 6.5×10^6cpm/1.25 pmole of the kinased synthetic oligonucleotide/2.5 ml/filter at 50.5°C, overnight, washed twice in 6X SSC, 40°C, 5 min., and once in 6X SSC, 50.5°C, 5 min. and autoradiographed overnight at -80°C.

Nucleotides 1-15 of the above synthetic oligonucleotide are
30 complementary to the 5'-end sequences of the CBHI gene, and nucleotides 23-37, to the control region between the S/D sequence and the adjacent BamHI site in ptrpCBH8. As seen in the expanded portion in Figure 11, the desired recombinant phage, in which the oligonucleotide primer has replaced an end-homologous ptrpCBH81

region, has lost BamHI and XmaI sites, and acquired ClaI and AhaIII sites derived from the central portion of the oligonucleotide. In particular the modified phage contains a TTTAAA AhaIII recognition site immediately adjacent the ATG start codon of the CBH gene, as indicated in the lower expanded portion of the figure.

5

Candidate plaques were picked and replication form (RF)-DNA analyzed, following infection of E. coli K12 strain DG98, for acquisiton of the new ClaI and AhaIII recognition sites and loss of an XmaI recognition site. One candidate showing the appropriate analysis was designated mp10w5A.

10

D.8.c.2. Replacement into ptrp CBH81

Each of the plasmids ptrpCBH81 and mp10w5A RF-DNA were digested to completion with SalI and BamHI, and the digest ligated under standard conditions. The ligation mixture was digested with XmaI to inactivate undesired ligation products and used to transform E. coli K12 MM294 to Amp^R. Candidate plasmids were screened by restriction analysis to release: a diagnostic 1.26 kb DNA fragment with SalI/BamHI digestion; a 1.07 kb DNA fragment by ClaI/BamHI digestion; and a 1.07 kb DNA fragment by BamHI/AhaIII digestion. The correct construction, designated ptrpCBH82 in Figure 11, was confirmed by sequencing.

15

20

9. Construction of Yeast Expression Plasmids Encoding CBHI

Several plasmids were constructed placing the CBHI coding sequences under the control of the ENO1 promoter, retaining the vector backbone features of the bacterial/yeast replicating plasmids of subparagraph D.8 derived from pDG151. The construction of one of these vectors, designated penoCBH500.202, is illustrated in Figure 12.

25

Plasmid' pPM14 was used as a source of the ENO1 promoter. pPM14 is a derivative of pBR322 containing a 722 bp EcoRI/HindIII fragment from peno46 (Holland, M. J., et al, J. Biol. Chem. (1981) 256:1385) which encompasses a fragment running from 723 bp preceding the enolase-I translation initiation site to nucleotide -2 of the

30

enolase ATG initiation codon. pPM14 was digested with EcoRI, treated with Klenow in the presence of dATP and dTTP, and then digested with HindIII to provide the EcoRI(blunt)/HindIII ENO1 promoter containing fragment.

5   Plasmid ptrpCBH81 was digested with SalI, treated with Klenow in the presence of all four dNTP's, and then digested to completion with HindIII.

     The fragments prepared in the previous paragraphs were ligated at an equimolar ratio under sticky end conditions, followed by 10   blunt end ligation, and the ligation products were treated with NruI and SmaI to inactivate unwanted ligation products. The ligation mixture was used to transform E. coli K12 MM294 to Amp^R and the transformants were screened for the presence of the desired 11.32 kb plasmid, designated penoCBH24. The correct construction was confirmed 15   by restriction analysis.

     With continued reference to Figure 12, ptrpCBH82 was digested with AhaIII and the 2.2 kb DNA fragment which encodes CBHI was purified using agarose gel electrophoresis. The two AhaIII sites cleaved by this procedure are (1) the AhaIII site introduced adjacent 20   the 5' end of the CBHI gene, as just described and (2) an AhaIII site contained in the pDG151 vector region which borders the 3' end of the CBHI gene in ptrpCBH81 and ptrpCBH82. This fragment was then digested with BamHI yielding two subfragments, a 1068 bp AhaIII/BamHI fragment encoding the 5' CBHI sequences and a 1.13 kb BamHI/AhaIII fragment 25   encoding the 3' CBHI sequences, and which further contains a stretch of bacterial DNA and the 3' untranslated ENO1 sequences.

     Plasmid penoCBH24 was digested with HindIII, treated with S1 nuclease, and then digested with BamHI. The foregoing fragments and the vector DNA fragment from penoCBH24 were ligated at a 2:1 molar 30   ratio under sticky end conditions, then under blunt end conditions. The ligation mixture was used to transform E. coli K12 strain MM294 to Amp^R and successful transformants screened for the desired 12.2 kb plasmids containing 5' CBHI sequences. Successful candidates, which were identified by release of a 2.05 kb fragment upon digestion with

EcoRI, and a 200 bp fragment upon digestion with HaeIII, were designated penoCBH500.X, where X is an indicator number specified below. The plasmids were further analyzed by dideoxy DNA sequence analysis using an internal octadecamer primer which is complementary to nucleotides 31-48 of the mature CBHI coding sequence.

Plasmid penoCBH500.202 encoded the sequence:

5' ...CTAAATCAA AAAATG TAT CGG ...3'
ENO1 CBHI

indicating a precise elimination of the HindIII cohesive end and fusion to the AhaIII/CBHI DNA fragment.

Plasmids penoCBH500.108 and penoCBH500.150 encoded identical DNA sequences:

5' ...CTAAATCA AAAATG TAT CGG ...3'
ENO1 CBHI

indicating removal of one A/T base pair during the S1 nuclease treatment.

Plasmid penoCBH500.212 encoded the sequence:

5' ... CTAAAT AAAATG TAT CGG ...3'
ENO1 CBHI

indicating removal of two A/T base pairs and one C/G base pair from the ENO1 promoter fragment.

D.10. Expression of Cloned CBHI Gene in Yeast

Yeast strain S. cerevisiae S173-6B was transformed with penoCBH500.202 and selected for LEU$^+$ transformants. A successful transformant colony was cultured and grown in a 10 liter fermenter, using glucose as a carbon source on minimal medium. The culture was grown to an optical density (OD_{680}) of 10. The contents of the fermenter were filtered, and the liquid fraction was concentrated and dialyzed against water.

The approximately 10 l of medium obtained was reduced to 500 ml and further purified by applying the material to a DEAE sepharose column according to the protocol described in paragraph D.2. above.

Of 500 mg protein applied to the column, about 437 mg was recovered in the non-absorbed fraction, and about 100 mg, or 18%, was eluted in a single peak corresponding to CBHI.

10.c. Characterization of CBHI Recombinantly Produced in Yeast

5   The purified, secreted CBHI (recombinant CBHI) was characterized and compared with native T. reesei CBHI which was isolated according to paragraph D.2 above. The results are summarized in Table II below.

      Western blot analysis of recombinant and native CBHI was 10 performed according to the standard procedures. Lanes 1 and 3 in Figure 13 show the gel patterns of native and recombinant CBHI, respectively. As indicated in Table II, the blot analysis indicates that the apparent molecular weight of native CBHI is about 60K daltons and that of recombinant CBHI, between about 100K-200K daltons.

15   The two enzymes were treated with the glycosidase endo-H, which acts to deglycosylate proteins at N-linked glycosyl residues. The deglycosylated native and recombinant proteins were examined as above by Western blot analysis, with the results shown in lanes 2 and 4, respectively, of Figure 13. As seen both deglycosylated proteins 20 have roughly the same molecular weight, indicating that recombinant CBHI, unlike native CBHI, is highly glycosylated by large N- linked residues. Native CBHI by contrast, is known to contain mainly O-linked residues, as noted in Table II. The much greater apparent molecular weight of glycosylated recombinant CBHI is attributed to the 25 greater length of N-linked sugar residues, as compared with the 1-2 sugar residues per site with O-linked glycosylation in the native enzyme.

      The CBHI activity of the purified native and recombinant enzymes was assayed conventionally for formation of cellobiose from 30 phosphoric acid-swollen cellulose. As seen from Table II, native CBHI has a specific activity, based on a Lowry protein determination, of about 0.57 units/mg, as compared with 0.34 units/mg for recombinant CBHI. The higher (0.46 units/mg) specific activity value for

recombinant CBHI was calculated on the basis of the protein amino acid composition. As seen, recombinant CBHI produced according to the present invention has nearly the same specific activity as native CBHI from T. reesei.

5

TABLE II

Comparision of Native and Recombinant CBHI

Property	Native CBHI	Recombinant CBHI
Type	Glycoprotein	Glycoprotein
Glycosylation	Mainly O-linked	Mainly N-linked
Apparent MW	60,000	100-200,000
Sp. Activity	0.57 units/mg	0.34 (0.46) units/mg

10

To confirm the identity of the recombinant CBHI, the amino acid sequence of an amino-terminus fragment (residues 1-37) and an internal fragment (residues 225-251) were determined, using methods

15 described above in paragraph D.2. The two sequences were each identical in amino acid composition to corresponding fragments of native CBHI. Interestingly, the recombinant protein contained the same amino-terminal pyroglutamine as the mature native CBHI, indicating identical processing of the recombinant enzyme involving

20 (1) removal of the leader polypeptide and (2) cyclization of the amino-terminal glutamine residue.

In summary, the gene encoding a CBH cellulase has been cloned and expressed in yeast, and by use of suitable control sequences is adaptable for expression in a variety of hosts. The

25 availability of recombinant CBH offers the possibility of more efficient use of cellulose waste products and conversion of these into useful derivatives, including, ultimately, glucose and ethanol and for specific modification of cellulose materials.

Although the construction and expression of a CBH gene has

30 been illustrated with respect to a T. reesei gene expressed in yeast, it is understood that the techniques described herein are adaptable to the construction of other CBH genes e.g., CBHII from a variety of

fungal sources. The general applicability of the expression vector illustrated herein to expression of other proteins is discussed in paragraph B.3 above.

E. Preparation of Recombinant EGI

5   This section illustrates the construction of a recombinant EGI gene, and its expression in yeast. The exemplary EGI coding sequence was constructed as a HindIII cassette, and ligated into a yeast expression vector in operable linkage to the yeast enolase control sequences.

10 E.1. Preparation of mNRA Enriched for EGI and Synthesis of cDNA

         Poly-A RNA was obtained from cellulose-induced T. reesei strain L27, and fractionated by agarose electrophoresis substantially as described in paragraph D.3 above.

         Antibodies specific against EGI were prepared using purified

15 EGI according to standard techniques. Briefly, purified EGI obtained by the method disclosed by Shoemaker, S., et al., Biotechnology (October, 1983) 687 was mixed with Freund's adjuvant and 0.2-0.4 ml of the mixture injected intramuscularly into New Zealand white rabbits, repeating the injection every 10 days until antibody titers were

20 detectable by the double immunodiffusion method of Ouchterlony, O., Arkiv Kemi (1949) 1:43.

         To verify the correct mRNA fraction, a small amount of each RNA fraction was added to the cell-free protein-synthesizing system sold as a reticulocyte lysate/methionine L-(35S)-translation system by

25 New England Nuclear Company, Boston, MA. The presence or absence of EGI synthesized by the system was verified by SDS-PAGE of total protein synthesized and of protein from the translation system immunoprecipitated with, and then dissociated from, anti-EGI antibody.

30 E.2. Preparation of a cDNA Probe

         The mRNA fraction which, in a cell-free translation system, results in the synthesis of high levels of EGI, was selected and used

to generate a single stranded copy DNA, substantially as described in paragraph D.4 above. The single stranded DNA was size analyzed by gel electrophoresis; a major band corresponding to the expected cDNA size was observed.

5   E.3.   <u>Preparation of a Double Stranded cDNA Library</u>

The mRNA fraction prepared from paragraph E.1 was also used to create a cDNA library according to the method of Maniatis, et al., <u>Molecular Cloning-A Laboratory Manual</u> (1982) Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pp. 235-238. The thus obtained
10   double stranded cDNA was then cloned into a pUC-8 cloning vector according to standard procedures, and probed using a 4 kb HindIII genomic fragment (described above) known to encode EGI.

Three cDNA clones hybridizing to the probe -- pcEG10, pcEG9, and pcEG1 -- were selected for further study. They were analyzed by
15   restriction analysis and by sequencing of the cDNA region using the dideoxy method. pcEG10 was found to encode the 3' region of the sequence, and pcEG1, to encode the 5' portion.

E.4   <u>Preparation and Sequencing of the EGI Gene</u>

A genomic library from <u>T. reesei</u> strain L27 in lambda phage
20   was prepared as described by Shoemaker, S., et al., <u>Bio Technology</u> (October, 1983) pp. 691-696, incorporated herein by reference, and as further described in paragraph D.5 above. The lambda genomic library was screened as described in Shoemaker, S., et al., (supra) but using the single stranded cDNA prepared in paragraph E.2, as probe. One
25   fragment from a positively reacting phage was subcloned into pUC8 to give pUC8-h. This vector includes a 4 kb HindIII fragment spanning the entire EGI gene.

The coding sequence was localized within this fragment by restriction analysis and hybridization to 5'-end synthetic
30   oligonucleotides and 3'-end cDNA probes. The DNA was digested with KpnI, SstI, XbaI, and SphI in tandem with EcoRI, and subjected to Southern blot analysis (Southern, <u>J. Mol. Biol.</u> (1975) <u>98</u>:503), using

as a probe a 16mer corresponding to amino acid residues 16-20 of the mature EGI sequence, with 4 nucleotide ambiguities. The results indicated that a major portion of the gene was located between the KpnI and SstI site. A restriction map of the pUC8-h genomic DNA is shown in Figure 14, frame A.

The identity of a portion of the 4 kb HindIII sequence with the EGI coding sequence was further confirmed by a restriction/selection experiment. The 4 kb fragment was bound to a nitrocellulose filter according to the procedure of Harpold, M. H., et al., *Nucleic Acids Res.* (1978) 5:2039. Total mRNA from induced cultures was added to the filters and mRNA which bound under stringent conditions separated by washes from unbound mRNA. The filter-bound mRNA was then eluted. For each filter, bound and unbound mRNA fractions were translated in the cell free protein synthesizing system and the resulting synthesized proteins precipitated with a mixture of anti-EGI antibodies and *Staphylococcus aureus* cells. The protein from the immunoprecipitate was subjected to SDS-PAGE, and the results of the experiment are shown in Figure 15. Lane 1 in the figure shows some of the molecular weight markers which were run in parallel. Lane 2 is a gel of total poly A-RNA translation products produced in the cell-free synthesizing system. The anti-EGI immunoprecipitate of the total translation products is displayed in lane 3, showing a single band corresponding in molecular weight to EGI. The gel pattern of translation products of poly A-RNA which hybridized to the probe is seen in lane 4, which shows a major band corresponding to the molecular weight of EGI. The lane 4 material, after immunoprecipitation by anti-EGI antibody, gave the single band shown in lane 5 having the expected molecular weight for EGI. Lanes 6 and 7 show, respectively, are total and an anti-EGI immunoprecipitate of translation products directed by poly A-RNA fraction which did not hybridize the 4 kb HindIII fragment. As seen from the figure, only the mRNA that bound to the filter containing the 4 kb HindIII fragment was capable of directing the synthesis of protein which was immunoprecipitated by anti-EGI.

The entire coding sequence of the EGI gene was then obtained using standard restriction and sequencing techniques. The results are shown in Table III below. The sequenced fragment contains 1697 nucleotides, including 66 nucleotides for a 22 amino acid leader

5 sequence preceding the coding sequence for the 437 amino acid residues of the mature protein, (as deduced from the known N-terminal sequence of the secreted purified EGI.) The DNA sequence also contains two introns at nucleotide positions 893 through 962 and 1553 through 1609. The intron positions were deduced by comparison of the coding

10 sequences with those obtained by analogous sequencing of the cDNA plasmids pcEG10 and pcEG1, and by comparison with the known amino acid sequence (also shown in the table). The coding sequence corresponding to the leader peptide (amino acids 1-22) provides a signal sequence for initiating secretion of EGI, and is referred to herein as the EGI

15 signal sequence. The amino acid sequence in Table III, in contiguous form, is the polypeptide sequence of EGI.

TABLE III

```
GTTACTCTGT  GTTGCAAATC  CAACTTATAA  AGACACAAC  CGCAAACTTT  GTCTGTGTGGC  CATCGGATTG  TTGGCCAAGC  ACCCTCCCCC  CCCCCTATCT  TAGTCCTTCT  TGTTGTCCCAAA  ATG
                                                                                                                                                  MET

GGG CCC TCA   GTT ACA   CTG CCG ACC   TTG ATC CTG   GCC ATT GCC   GCC CTG GTC   GCC ACC CCC   GAG GTC CAT   CCC AAG
Ala Pro Ser   Val Thr   Leu Pro Thr   Leu Ile Leu   Ala Ile Ala   Ala Leu Val   Ala Thr Pro   Glu Val His   Pro Lys

TTG ACA ACC   TAC AAG   TGT AAG GGG   TCC GGG TGG   CGG TAC AAC   GTT GTC TGC   ATG TGG ATG   GAC CAC TAC   TAC AAC
Leu Thr Thr   Tyr Lys   Cys Lys Gly   Ser Gly Trp   Arg Tyr Asn   Val Val Cys   Met Trp Met   Asp His Tyr   Tyr Asn

TCG ACC AAC   GTC GGC   AAC GAG ATC   GGG GAG TTC   TGC GAC CTG   AAG GGC GTC   GGC ATC GAG   GCC TAC GCC   TGG GGC
Ser Thr Asn   Val Gly   Asn Glu Ile   Gly Glu Phe   Cys Asp Leu   Lys Gly Val   Gly Ile Glu   Ala Tyr Ala   Trp Gly

GTC ACC GGG   AGC TCG   TAC AGC AGC   TCT AGC CCT   GGC GGC TCT   GGC GGC GGC   TAC TCT CTG   CTC CTG GAC   TCT GGG
Val Thr Gly   Ser Ser   Tyr Ser Ser   Ser Ser Pro   Gly Gly Ser   Gly Gly Gly   Tyr Ser Leu   Leu Leu Asp   Ser Gly

GGT TAC ATG   AGC CTC   GAG GAG CTC   GTC GAC GTC   TGT CGG CTG   GGA GAG AAC   GGC TGG TCC   CTC TAC CTG   TCT CAG
Gly Tyr Met   Ser Leu   Glu Glu Leu   Val Asp Val   Cys Arg Leu   Gly Glu Asn   Gly Trp Ser   Leu Tyr Leu   Ser Gln

ATG GAC AAC   GGG GCC   TAT CAG AAC   GGT TAC AAC   TGC GAT GCC   TAC GGC AGC   CCC GTC GTC   AOA TGG AGG   AAC GGC
Met Asp Asn   Gly Ala   Tyr Gln Asn   Gly Tyr Asn   Cys Asp Ala   Tyr Gly Ser   Pro Val Val   Thr Trp Arg   Asn Gly

ACC CTC AAC   ACT CAG   CAC AGC AAC   GGC TTC TGC   AAT GAG ATC   CTG AAC AAT   GGG AAT GCC   CCT ACC TGC   ACG GCC
Thr Leu Asn   Thr Gln   His Ser Asn   Gly Phe Cys   Asn Glu Ile   Leu Asn Asn   Gly Asn Ala   Pro Thr Cys   Thr Ala
```

```
GCC TGC GAC TCT GCC GGT TGC GGC TTC AAC CCC TAT GGC AGC GGC TAC AAA AGg tgagcctgat gccactacta cccctttcct gggcgtctcg cggtttccca tgctgacat
Ala Cys Asp Ser Ala Gly Cys Gly Phe Asn Pro Tyr Gly Ser Gly Tyr Lys Ser

ggtttccagC TAC TAC ATC ACC TTC ATC ACC CAG TTC AAC AGG GAC AAC GGC CCC TGG GGC AAC CTT
           Tyr Tyr Ile Thr Phe Ile Thr Gln Phe Asn Arg Asp Asn Gly Pro Ser Gly Asn Leu

GTG AGC ATC ACC GAC GTT CAA AAC GGC GGG CCC GAC ACC TGC TCC GCC TCA GCC TAC GGC
Val Ser Ile Thr Asp Val Gln Asn Gly Gly Pro Asp Thr Cys Ser Ala Ser Ala Tyr Gly

GGC CTC GCC AAG ATG AGC AGC GCC CTG GTG GGC ATG TTC AGC CTC GAC GAC AGC GGC AAC
Gly Leu Ala Lys MET Ser Ser Ala Leu Val Gly MET Phe Ser Leu Asp Asp Ser Gly Asn

GCC GGC CCC TGC AGC ACC GAG AAC GGC CCA TCC CAC ATC TTC AAC AAC CAC ATC TGG GGA GAC ATT GGG
Ala Gly Pro Cys Ser Thr Glu Asn Gly Pro Ser His Ile Phe Asn Asn His Ile Arg Trp Gly Asp Ile Gly

TCT ACT AAC TGG AGC CCC CCG TCC AGG AGG TTT ACT ACA AGG TGG ACT TGG AGC CGG AGC TGC AGG
Ser Thr Asn Trp Ser Pro Pro Ser Arg Arg Phe Thr Thr Arg Trp Thr Ser Ser Pro Ser Cys Thr

CAG ACT CAC TGG CAG CAG TAC GGC ATT TGC CAG AGG TGC TAT AGC AAC GAC Tgtccgatc cccatgcct
Gln Thr His Trp Gln Gln Tyr Gly Ile Cys Gln Arg Cys Tyr Ser Asn Asp

gacgggagt gatttgaya tgctaaccgc taaaataca gAC TCG TAC CAA TGC CTT TAGAGGGTTG ACTTGCCTCT GGTCTGTCCA GAGGGGGGGAC GATAGAATGC GGGCAGCAG GGAGCTGGTA
                                         Tyr Ser Gln Cys Leu
```

E.5. Construction of an Intronless EGI Coding Sequence

Plasmid pUC8-h from paragraph E.4 was digested with SphI, and the SphI fragment was purified on agarose gel. The isolated fragment was cloned into SphI-digested M13mp8 according to standard
5 methods. The ligated phage were transformed into competent E. coli K12 strain DG98 and cells plated in the presence of IPTG (5×10^{-4} M) on X-gal (40 g/ml) containing plates. Non α-complementing white plaques were screened for recombinant single strand phage DNA of the expected 10.3 kb size. The structure of the desired recombinant
10 phage, designated JV82.2, was confirmed using restriction analysis. This construction is shown in frame B of Figure 14, where the vector is designated Φ82.2.

The downstream intron was excised by replacing the 3' portion of the genomic clone with a corresponding portion of a cDNA
15 clone prepared from the enriched mRNA, and characterized by restriction analysis. Specifically, JV82.2 was digested to completion with PstI, and the digest ligated with a PstI digested fragment isolated from pcEG10, to obtain JV104.34,(Φ104.34) illustrated in frame C of Figure 14. The PstI fragment from pcEG10 spans the coding
20 region corresponding to that in the genomic clone which includes the portion containing the downstream intron. That is, JV104.34 contains the entire coding sequence of the EGI gene with the downstream intron removed.

To remove the upstream intron, site-directed mutagenesis was
25 employed. An oligonucleotide primer complementary to the sequences framing the first upstream intron, but lacking the intron sequences was synthesized. The primer had the sequence 5'-GGCAGCGGCTACCAAAAGCT-ACTACGGCCCCGGAGA-3'. The recombinant M13mp19 bacteriophage JV104.34 was prepared in E. coli K12 strain DG98 and the single strand phage
30 was purified conventionally. The single strand phage was used as a template for DNA synthesis to effect site-specific mutagenesis using the synthesized oligomer as primer, according to standard procedures. The successful bacteriophage candidate, JV129.3, (Φ129.3) obtained by screening with radiolabelled oligomer as probe, is

illustrated in frame D of Figure 14. JV129.3 contains the entire uninterrupted EGI coding sequence in proper reading frame.

The intronless gene was further refined, as shown in frames E and F of Figure 14, to provide HindIII sites immediately upstream

5   and downstream of the coding sequence thus permitting its manipulation as a HindIII cassette. To supply the upstream HindIII site, the synthetic oligonucleotide 5'-CTTAGTCCTTCTTGAAGCTTTAAAATGGCGCCCT-3', which is complementary to the sequence immediately upstream of the coding sequence, but with 6 mismatches to create a HindIII site

10 immediately 5' of the ATG translation initiation codon was used to replace the native upstream sequences by site-specific mutagenesis as above. The resulting phage was further treated with HindIII and religated to delete the portion of DNA between the new 5' proximal HindIII and the upstream site shown in Figure 14, frames A-D yielding

15 JV134.1 (Φ134.1 in Figure 14, frame E).

.   Similarly, the synthetic nucleotide 5'-AATGCC-TTTAGAAGCTTGACTTGCCT-3' was used in the site-specific mutagenesis procedure to create the M13 vector containing the EGI gene as a HindIII cassette, designated JV139.2 (Φ139.2 in Figure 14, frame F).

20 E.6.   Construction of Expression Vectors Containing the EGI Coding
        Sequence

An expression vector useful in yeast containing the EGI sequences under the control of the enolase-1 promoter and enolase-1 terminator was constructed by ligating the HindIII cassette into the

25 HindIII site of the yeast host vector pMAC101.

pMAC101 is a 10.2 kb vector containing a HindIII cleavage site disposed between a modified yeast enolase-1 (ENO1) promoter sequence and a yeast endolase-1 (ENO1) terminator sequence. The vector also contains the yeast LEU2 gene, a yeast 2 μ origin of

30 replication and the origin of replication and Amp^R portions of pBR322. pMAC101 can be obtained readily from vector pJV160, described below, by removing a HindIII gene cassette in that vector.

To construct the expression vector for EGI in yeast, designated pJV160, pMAC101 was digested to completion with HindIII, and ligated to the HindIII cassette excised from the replicative form (RF) of JV139.2. The ligation mixture was used to transform yeast

5 strain C468. The successful transformants were selected for LEU$^+$, and positive candidates were assayed by their ability to clear an overlay of carboxymethyl-cellulose (CMC) as described below. (Of the cellulase-complex enzymes, only endoglucanases have the appropriate specificity to hydolyze carboxymethyl cellulose. Cellobiohydrolases

10 are specific for cellulose per se and fail to hydrolyze CMC.)

E.7. Confirmation of Expression of EGI in Yeast

After selection for LEU2$^+$, the three successful transformants were assayed for EGI formation, substantially according to the method of Teather, R., et al., Appld. Envir. Micro. (1982)

15 43:777. Colonies were grown on minimal plates at 30°C, and overlayed with top agar (0.8%) containing 0.5% CMC, 15 mM NaOAc, pH 4.5 and 10 mM sodium azide. The overlay plates were then incubated at 37°C for 4 hrs. To visualize clearing of the CMC, the plates were treated for 15 min. with 1 mg/ml Congo Red and for 5 min with 1 M NaCl followed by 5

20 min. with 1 M HCl. To test the ability of the three yeast transformants to produce secreted, functionally active EGI, a yeast culture supernatant from each of the three transformants, as well as from a yeast transformed by pMAC101 (control) was prepared. The supernatant samples (either 1x or 20x concentrated) were spotted on

25 CMC plates and assayed for EGI activity as described above. The results are shown in Figure 16. Spots 1-4 are 1x concentrates of the three EGI (1-3) and control (4) transformants, and spots 5-8 are the corresponding 20x concentrates. As seen, all of the EGI transformants, but not the control transformant, secrete active EGI.

30 Spot 9 was produced by native EGI from T. reesei.

Verification of the expression of EGI was also made by Western blot analysis. Concentrated culture supernatants from three yeast cultures which showed EGI activity by the previous assay were run on SDS-PAGE, transferrred to nitrocellulose, and probed with anti-

EGI antibody, followed by staph A protein radiolabelled with ^{125}I.

Lane 1 in Figure 17 shows native, purified EGI from T. reesei. Lanes 2, 3, 4 are gel patterns of the supernatant from each of the three EGI yeast transformants. As seen, each of the EGI transformants secretes a polypeptide having a molecular weight somewhat greater than that of native EGI. This band was not seen in the supernatant from from yeast transformed with pMAC101. The greater molecular weight of recombinant EGI is presumably due to differences from native EGI in post translational processing, similar to what has been described for CBHI.

Techniques used for the successful expression of EGI in yeast, described above, are applicable, by suitable modifications in vector control sequences, for achieving expression in a variety of hosts, including both bacterial and yeast hosts. As with CBHI, the methods are applicable to cloning and obtaining expression of a variety of other endoglucanases, including EGII from T. reesei.

Particularly in combination with recombinant CBHI, the availability of recombinant EGI offers the possibility of more efficient use of cellulose products and conversion of these into useful derivatives, and/or selective modification of cellulosic materials. A cellulose degrading system containing CBHI, EGI and optionally, β-glucosidase, may be constructed from individually isolated recombinant enzymes or by using a multi-host system which contains the combination of cellulases.

In addition to the ATCC and NRRL deposits identified above, the following materials have been have been deposited at the American Type Culture Collection, 12301 Parklawn Drive, Rockville, MD, U.S.A. (ATCC) under the terms of the Budapest Treaty on the International Recognition of the Deposit of Microorganisms for the Purposes of Patent Procedure and Regulations thereunder (Budapest Treaty) and are thus maintained and made available according to the terms of the Budapest Treaty. Availability of such strains is not to be construed as a license to practice the invention in contravention of the rights granted under the authority of any government in accordance with its

patent laws.

The deposits have been assigned the indicated ATCC deposit numbers, and have also been deposited with the Master Culture Collection (CMCC) of Cetus Corporation, Emeryville, CA, U.S.A., the assignee of the present application, and assigned the indicated CMCC deposit numbers:

Deposit or Host	CMCC#	Date of Deposit	ATCC#
DG98	1965	July 13, 1984	39768
pJV160	2068	July 13, 1984	39766
penoCBH500.202	2062	July 13, 1984	39767

WHAT IS CLAIMED IS:

1. A recombinant DNA sequence which encodes, without interruption, an amino acid sequence substantially equivalent to the amino acid sequence of a cellulase derived from a species of fungus.

2. The DNA sequence of Claim 1, wherein the species of fungus is a species of. Trichoderma.

3. The DNA sequence of Claim 2, wherein the species of fungus is T. reesei.

4. The DNA sequence of Claim 1, wherein the cellulase is a cellobiohydrolase.

5. The DNA sequence of Claim 4, wherein the cellobiohydrolase is T. reesei cellobiohydrolase-I.

6. The DNA sequence of Claim 1, wherein the cellulase is an endoglucanase.

7. The DNA sequence of Claim 6, wherein the endoglucanase is T. reesei endoglucanase-I.

8. The DNA sequence of Claim 1, wherein the cellulase is T. reesei β-glucosidase.

9. A mature, glycosylated, functionally active fungal cellulase produced by yeast transformed with a yeast expression vector containing a fungal-derived cellulase gene.

10. The cellulase of Claim 9, which is substantially equivalent, in amino acid sequence, to a Trichodermal cellobiohydrolase.

11. The cellulase of Claim 10, which is substantially equivalent, in amino acid sequence, to cellobiohydrolase-I from T. reesei.

12. The endoglucanase of Claim 7, which is substantially equivalent, in amino acid sequence, to a Trichodermal endoglucanase.

13. The cellobiohydrolase of Claim 12, which is substantially equivalent, in amino acid sequence, to endoglucanase-I from T. reesei.

14. A method of treating cellulose comprising:

providing an expression vector capable of transforming yeast to express a mature, glycosylated, functionally active fungal cellulase,

transforming yeast with such expression vector to produce such cellulase, and

reacting the cellulose enzymatically with the cellulase so produced.

15. The method of Claim 10, wherein the transformed yeast are capable of secreting the cellulase.

16. The method of Claim 15, wherein the cellulase produced is encoded by an uninterrupted coding sequence derived from a Trichodermal endoglucanase gene.

17. The method of Claim 16, wherein the cellulase produced is encoded by an uninterrupted coding sequence derived from a T. reesei endoglucanase-I gene.

18. The method of Claim 15, wherein the cellulase produced is encoded by an uninterrupted coding sequence derived from a Trichodermal cellobiohydrolase gene.

19. The method of Claim 18, wherein the cellulase produced is encoded by an uninterrupted coding sequence derived from a T. reesei cellobiohydrolase-I gene.

20. The method of Claim 15, wherein said reacting is carried out in the presence of such yeast.

21. The method of Claim 15, which further includes separating such cellulase from the yeast prior to said reacting.

22. The method of Claim 15, wherein such cellulase is a cellobiohydrolase and said reacting further includes reacting the cellulose with a β-glucosidase and/or an endoglucanase.

23. The method of Claim 15, wherein such cellulase is an endoglucanase and said reacting further includes reacting the cellulose with a cellobiohydrolase and/or a β-glucosidase.

24. An expression vector for expressing a mature heterologous protein in yeast comprising, in a 5'- to -3' direction,

a yeast promoter sequence,

a cellulase signal sequence operatively linked to the promoter sequences, and

the coding region of the heterologous protein operatively linked to the signal sequence.

25. The vector of Claim 24, which further includes a yeast terminator sequence operatively joined to the 3' end of said coding region.

26. The vector of Claim 24, wherein said promoter sequence includes a yeast enolase-1 promoter.

27. The vector of Claim 24, wherein the signal sequence includes the leader coding sequence derived from the T. reesei endoglucanase-I gene.

28. The vector of Claim 24, wherein the signal sequence includes the leader coding sequence derived from the T. reesei cellobiohydrolase-I gene.

29. The vector of Claim 24, wherein the coding region includes an uninterrupted coding sequence which encodes a fungal cellulase.

30. The vector of Claim 29, wherein the fungal cellulase is T. reesei endoglucanase-I.

31. The vector of Claim 29, wherein the fungal cellulase is T. reesei cellobiohydrolase-I.
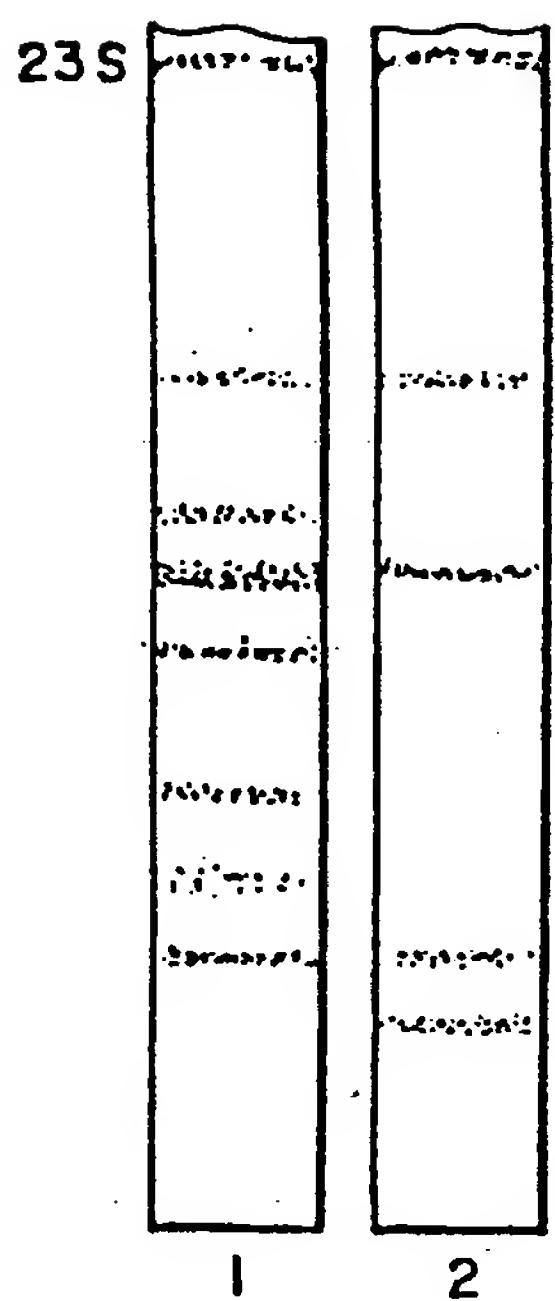
FIG. 1
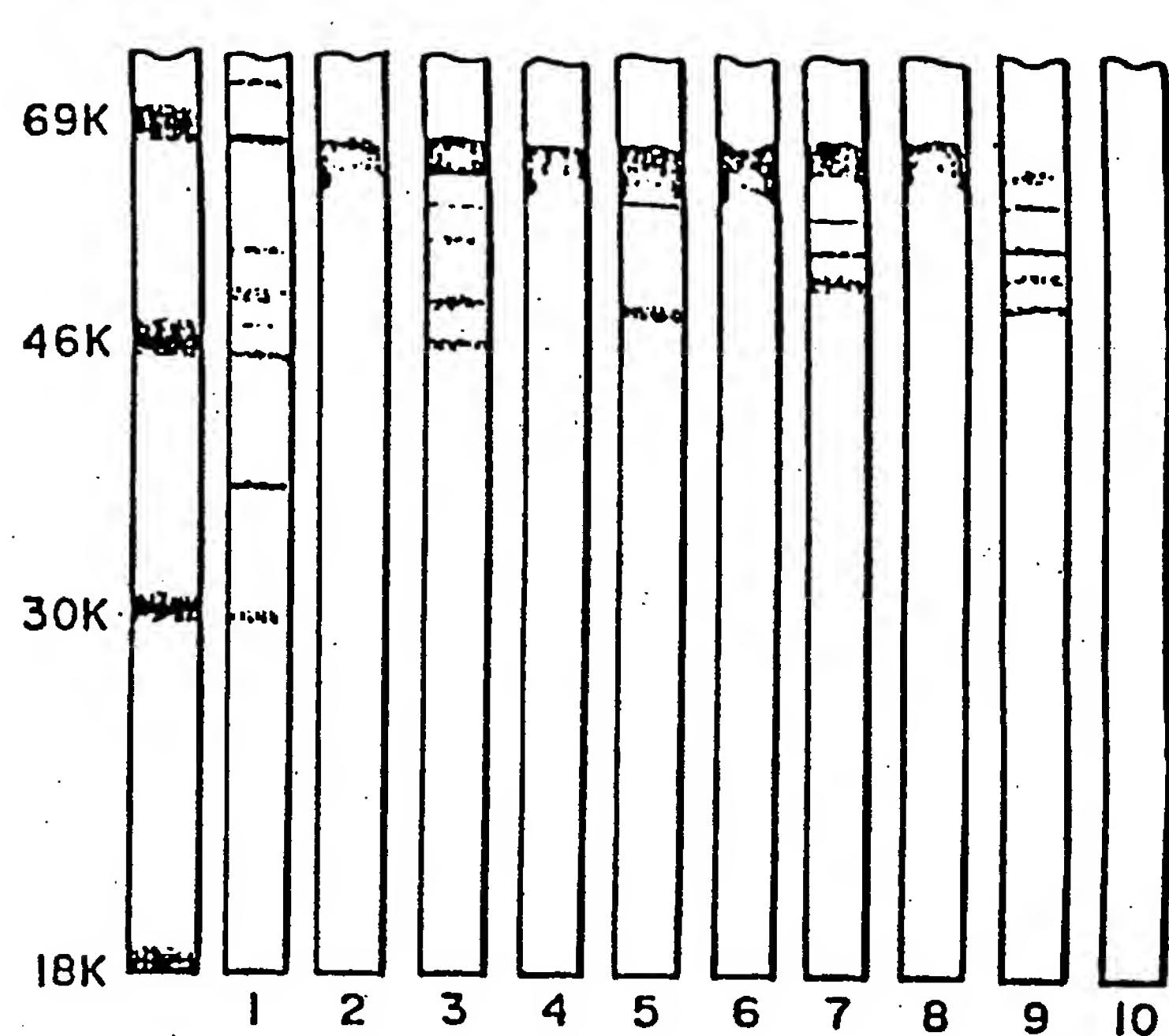
FIG. 2
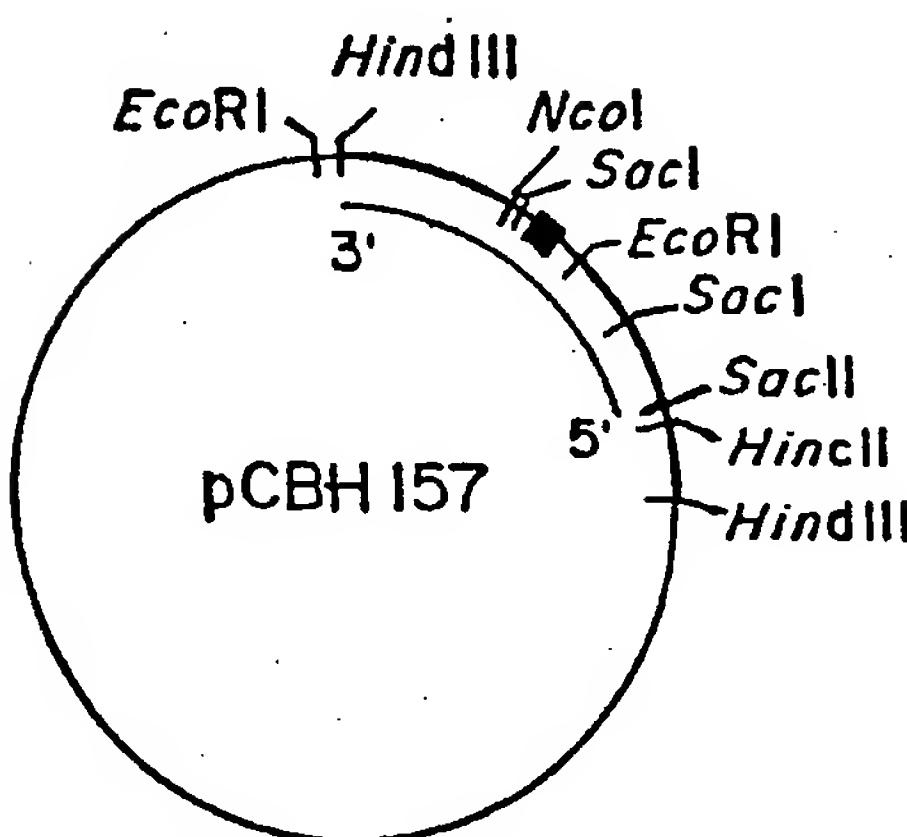
FIG. 3A
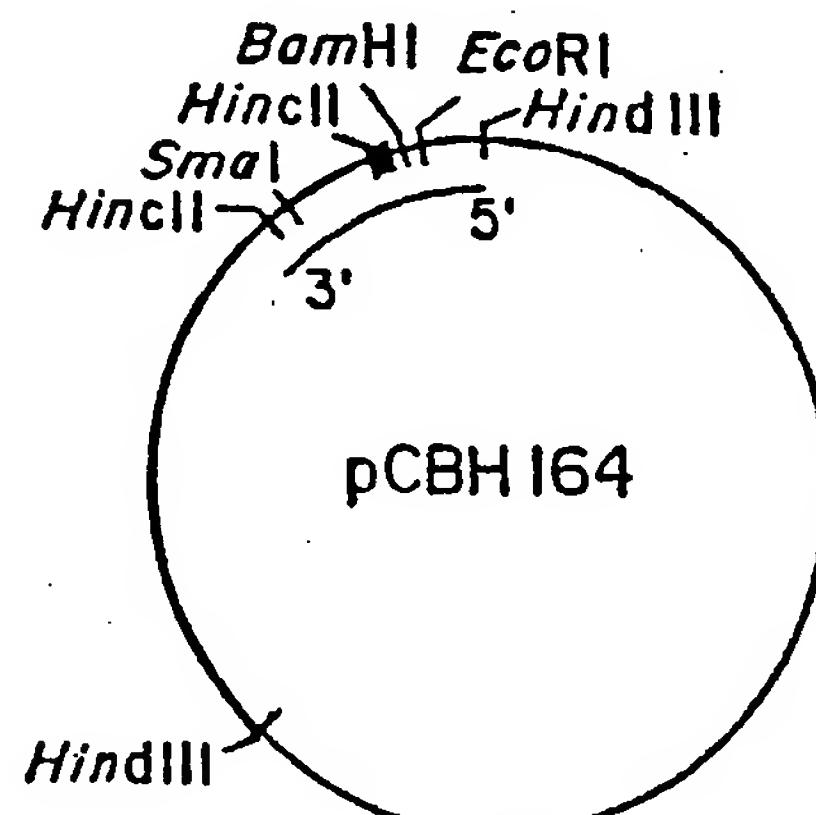
FIG. 3B
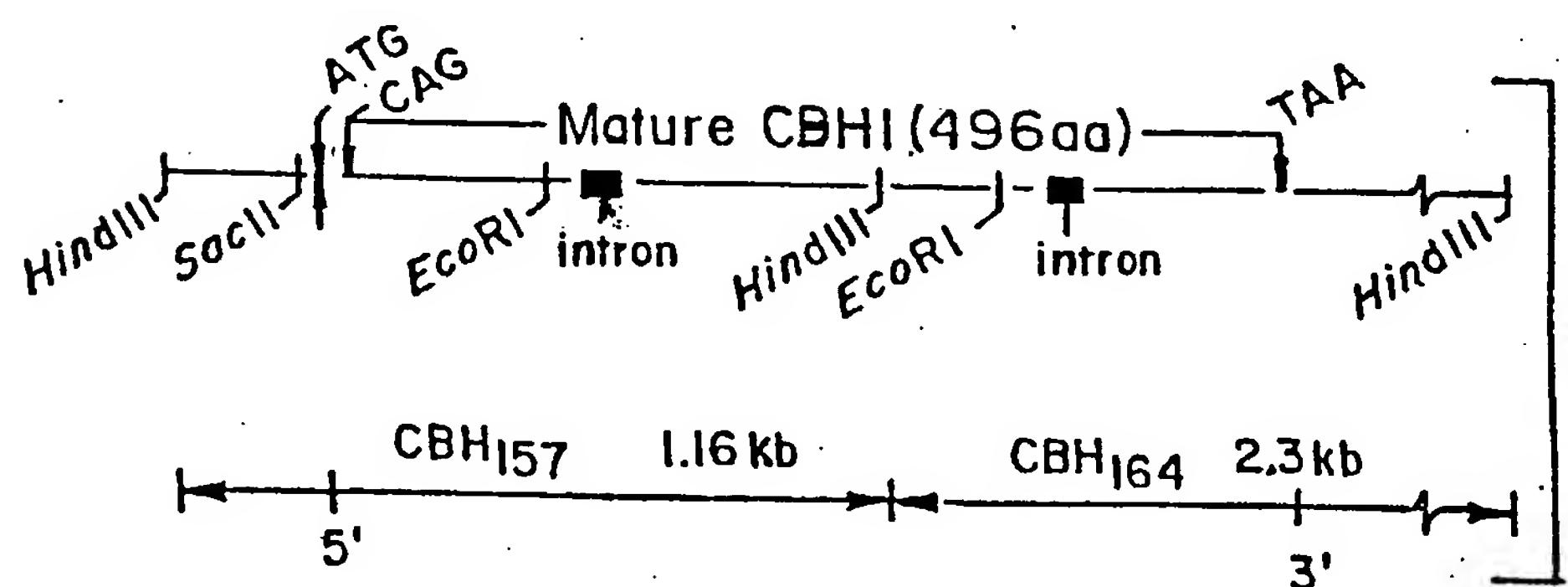
FIG. 4

FIG. 5
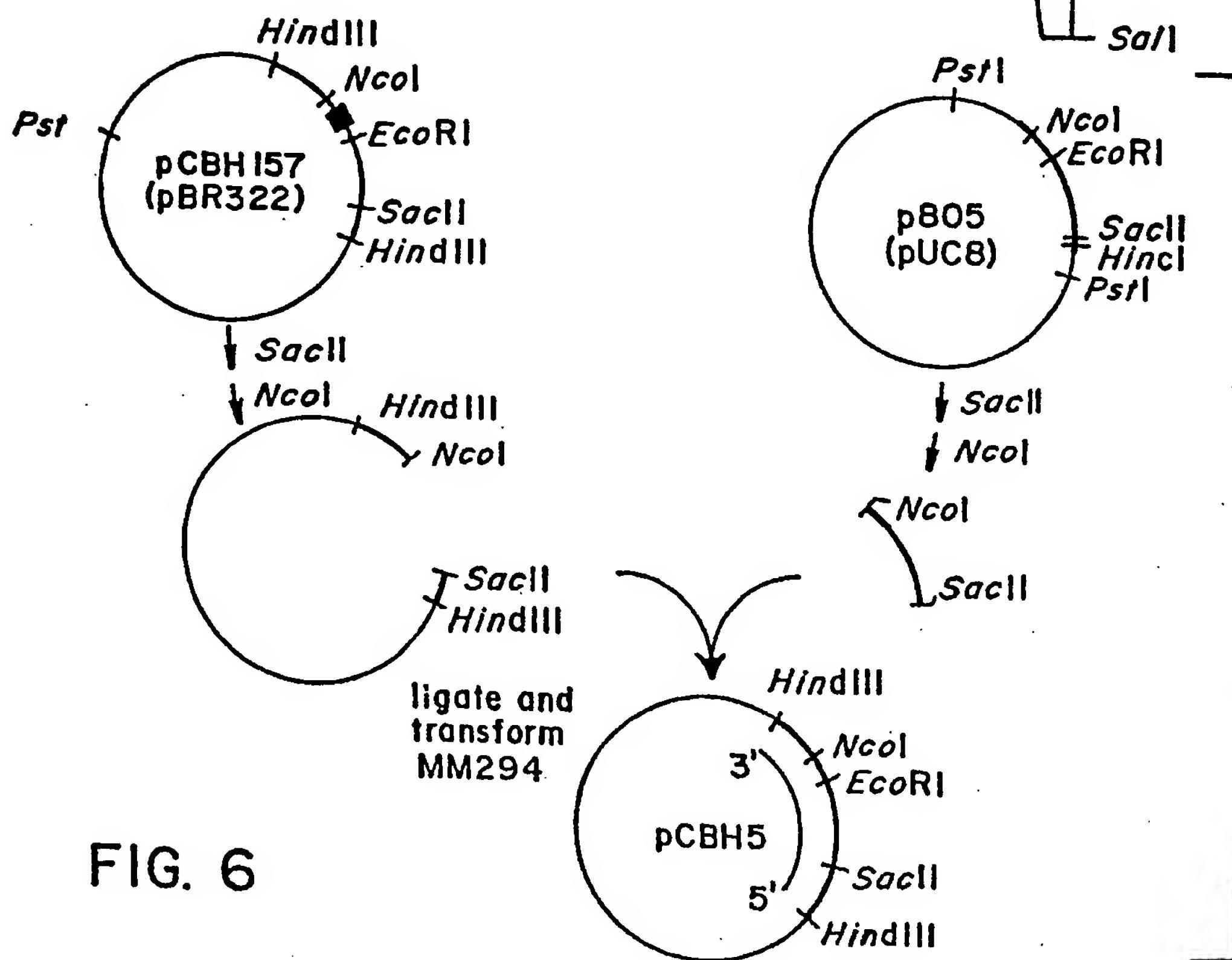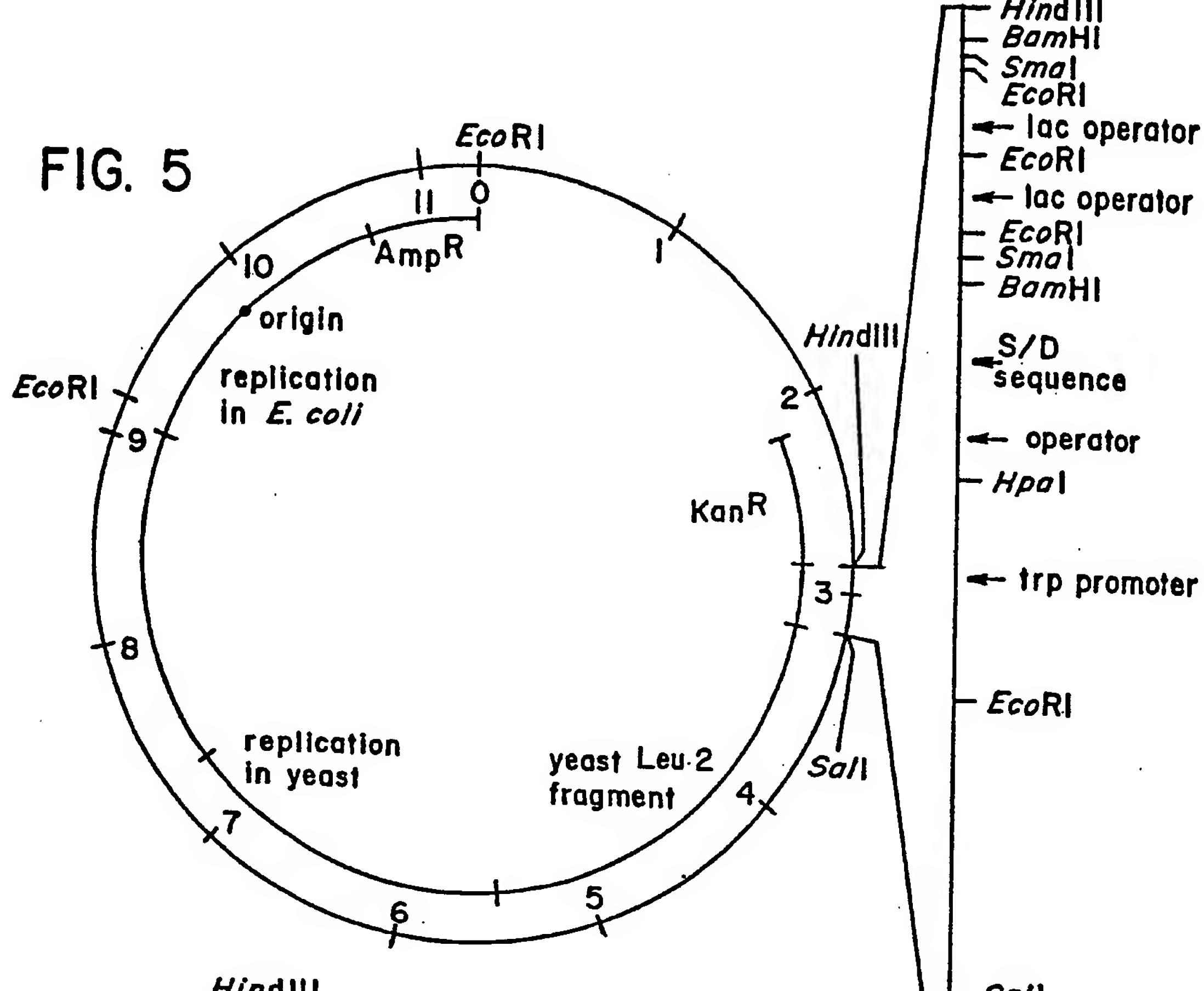
FIG. 6
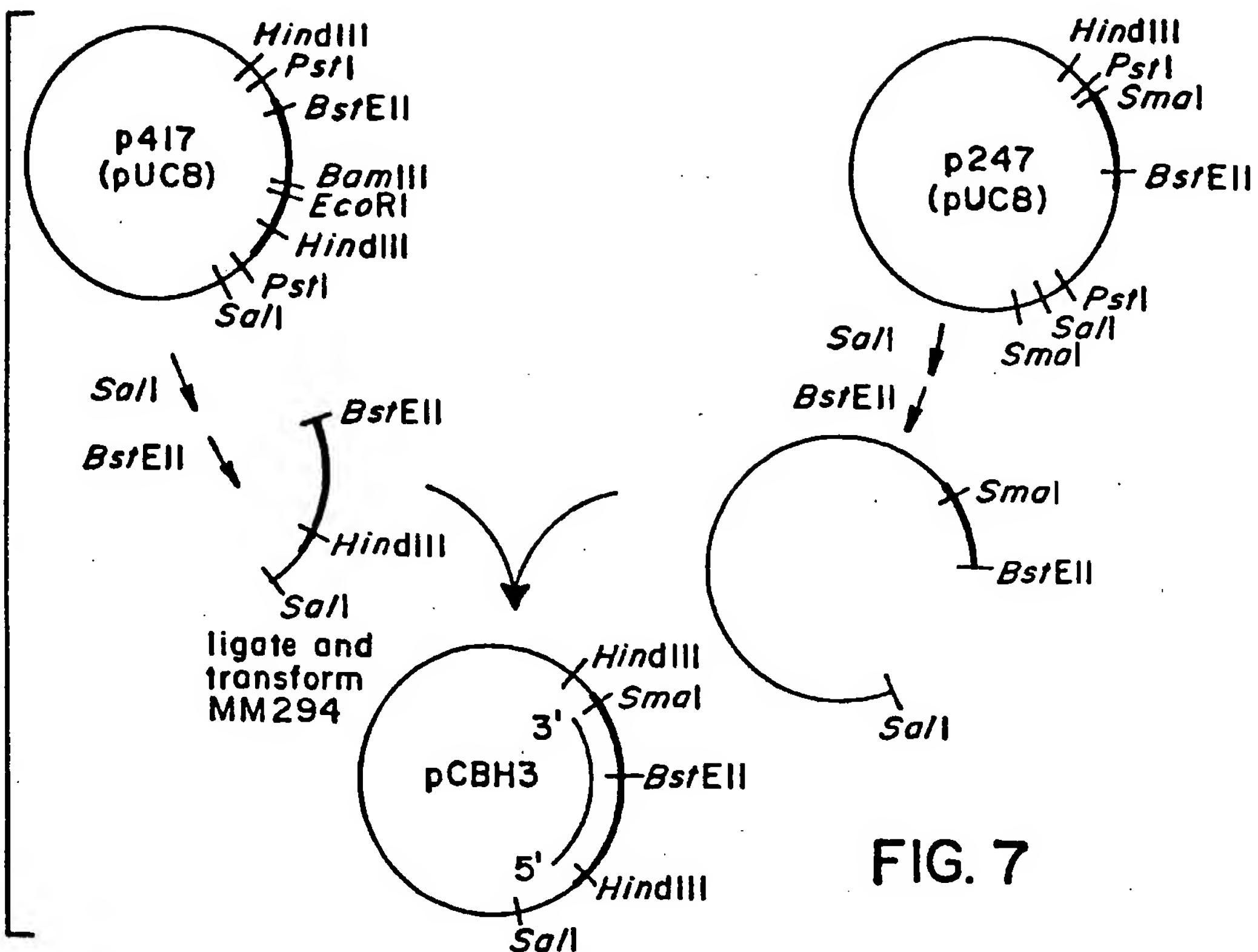
FIG. 7
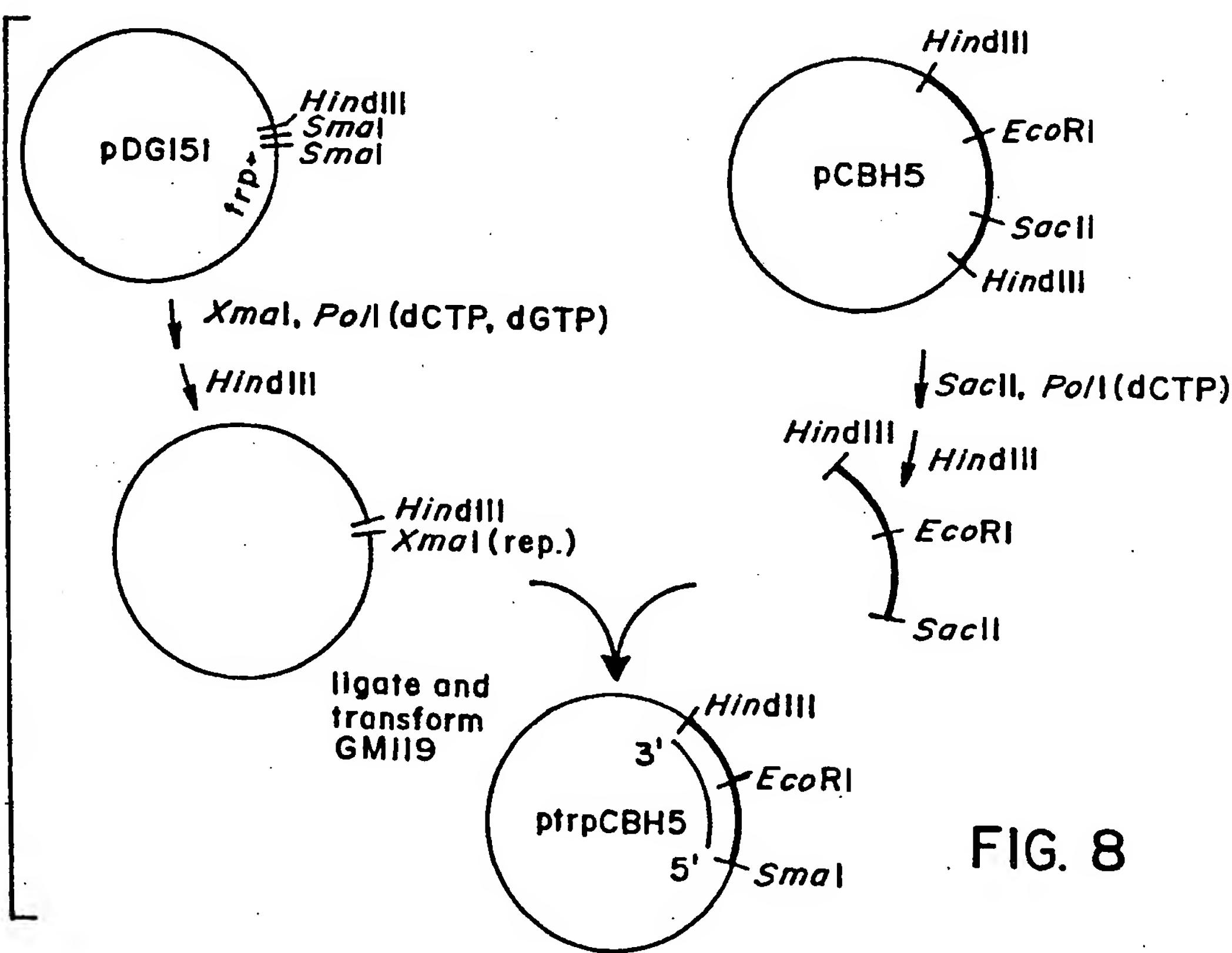
FIG. 8

0137280
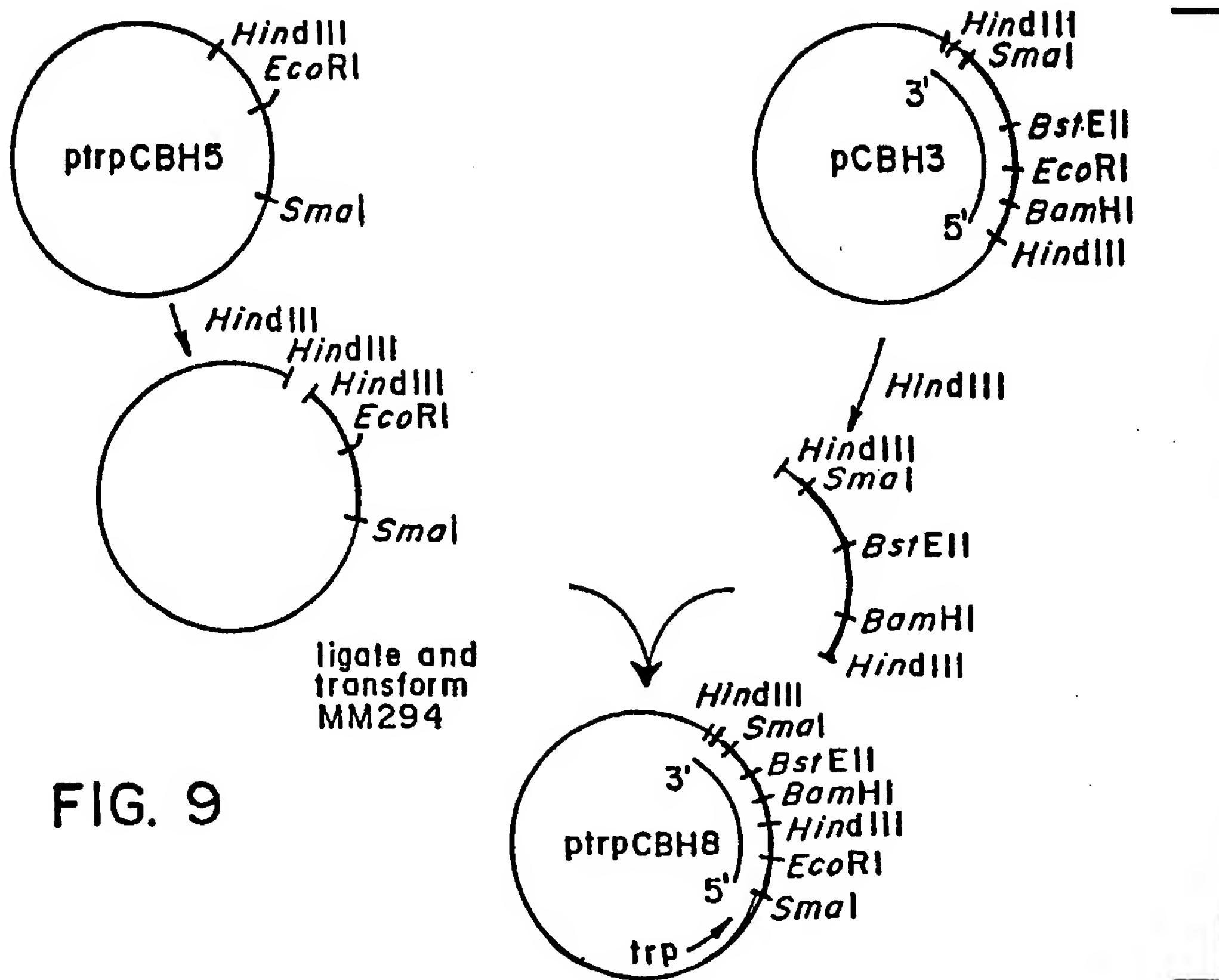
FIG. 9
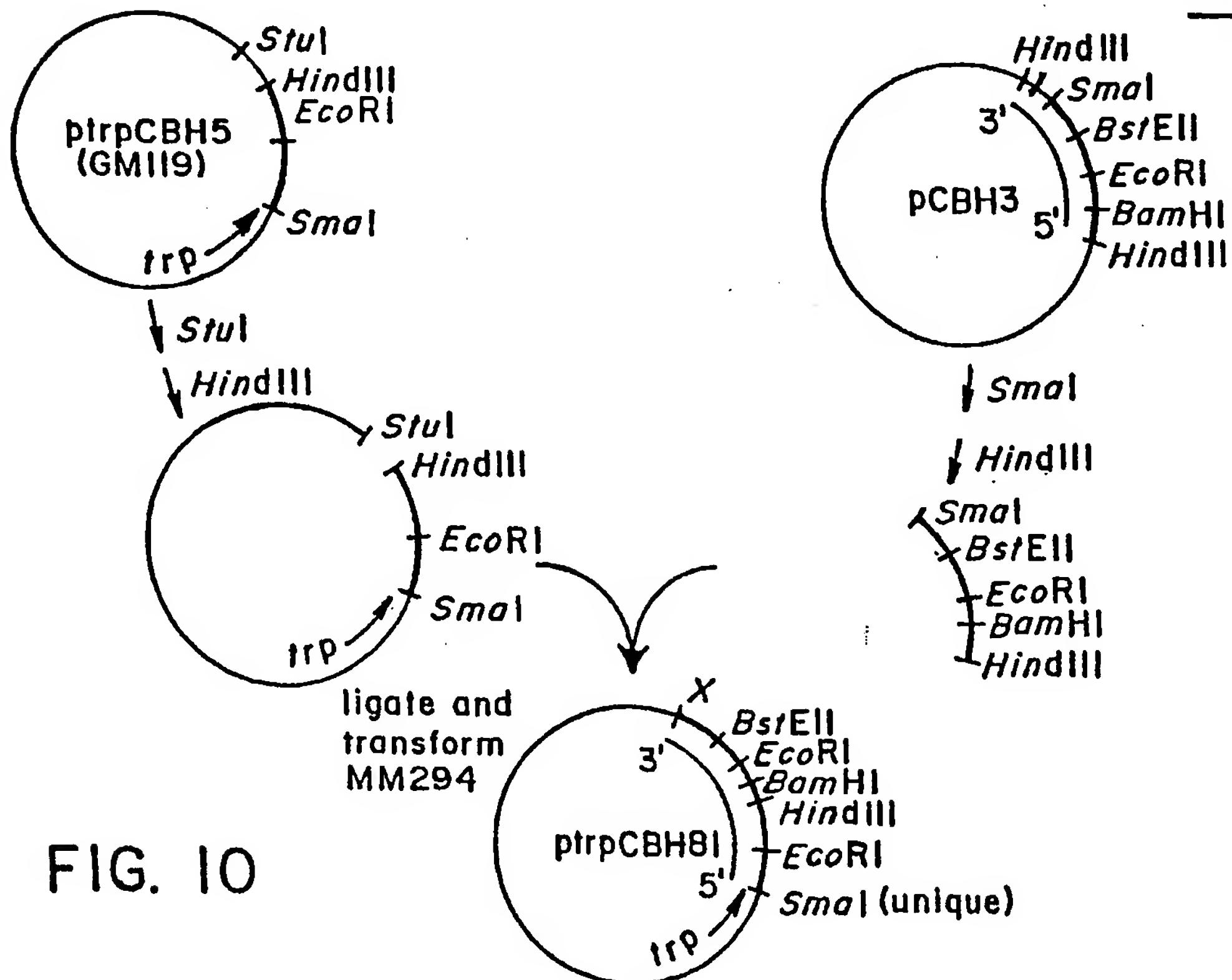
FIG. 10

0137280

FIG. 11

FIG. 12

FIG. 13

69K—
EGI—
45K—

30K—

18K—
12K—
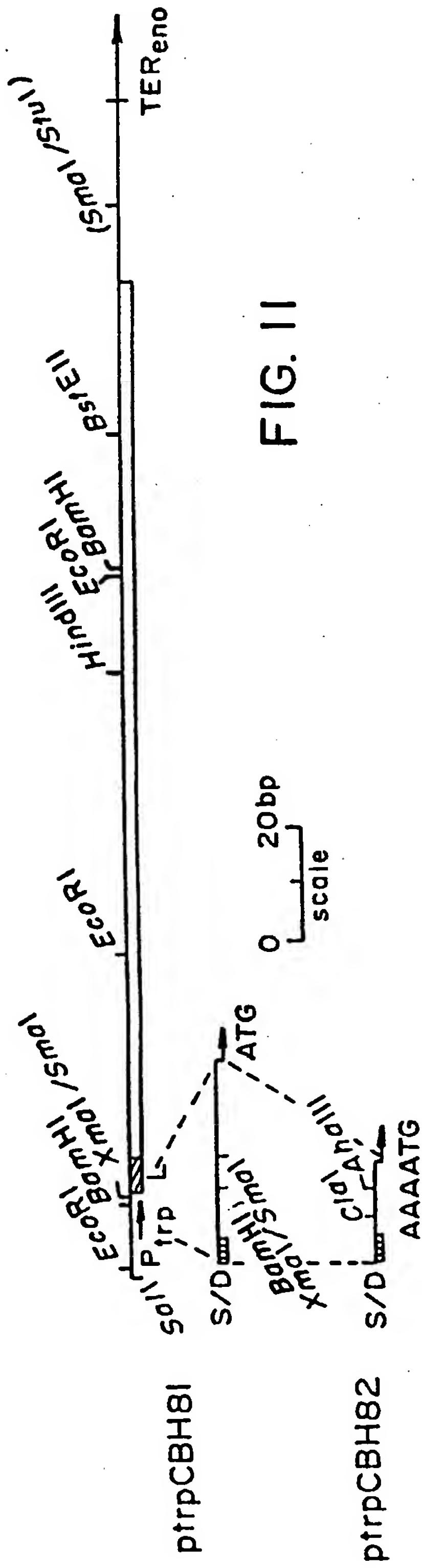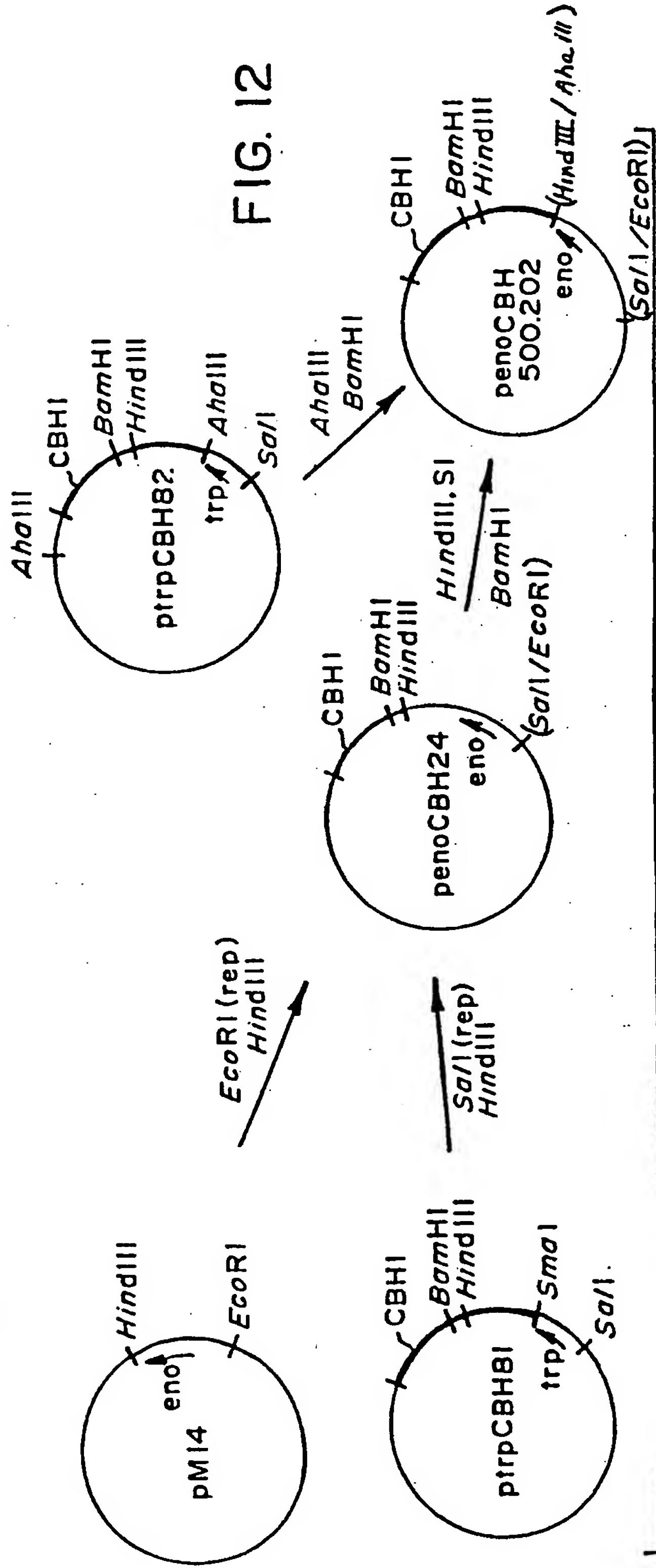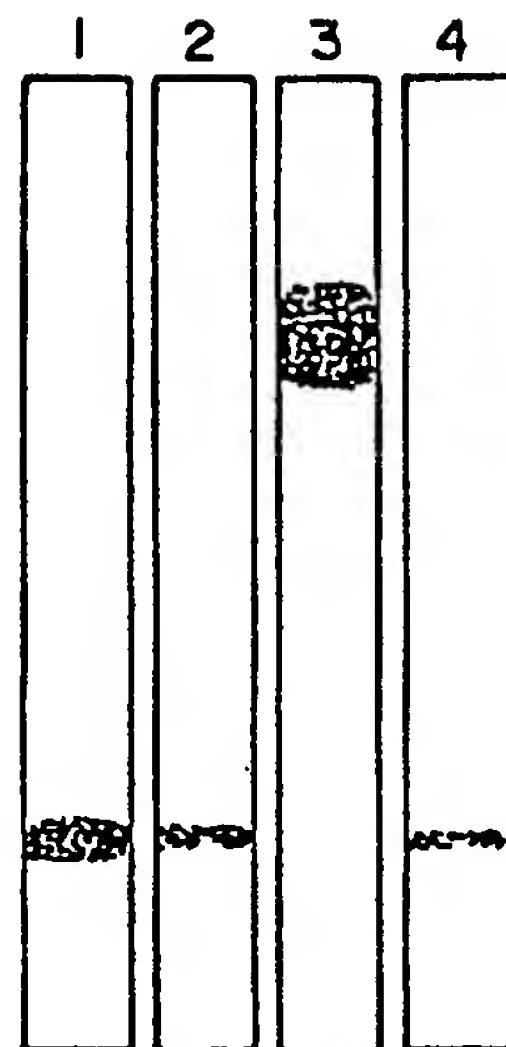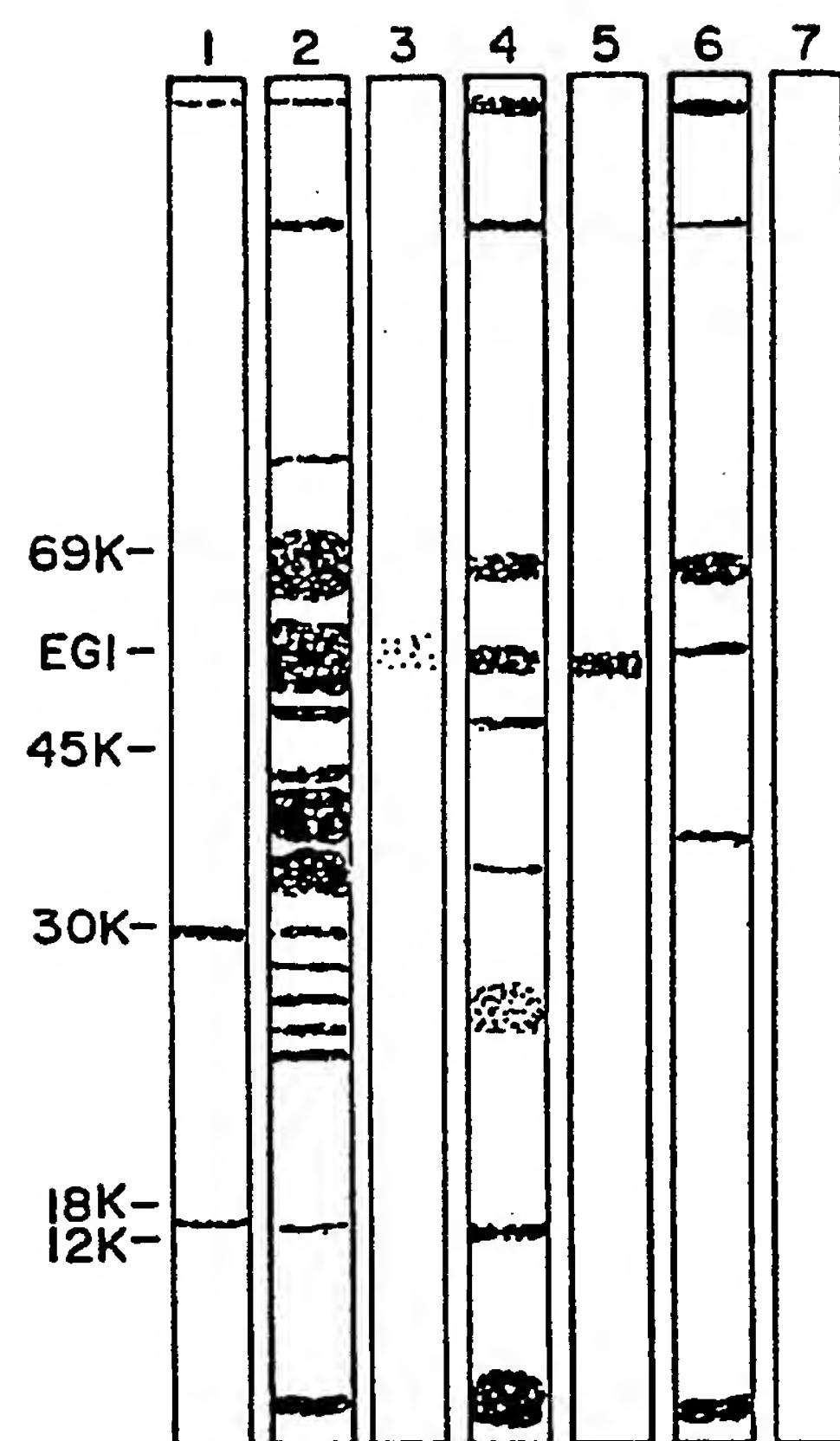
FIG. 15
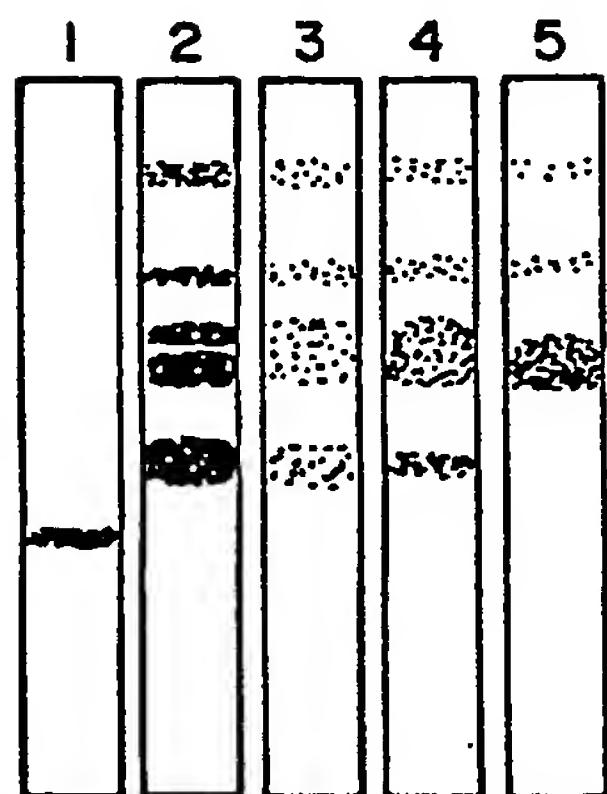
FIG. 17

—Recombinant EGI
—Native EGI

FIG. 16
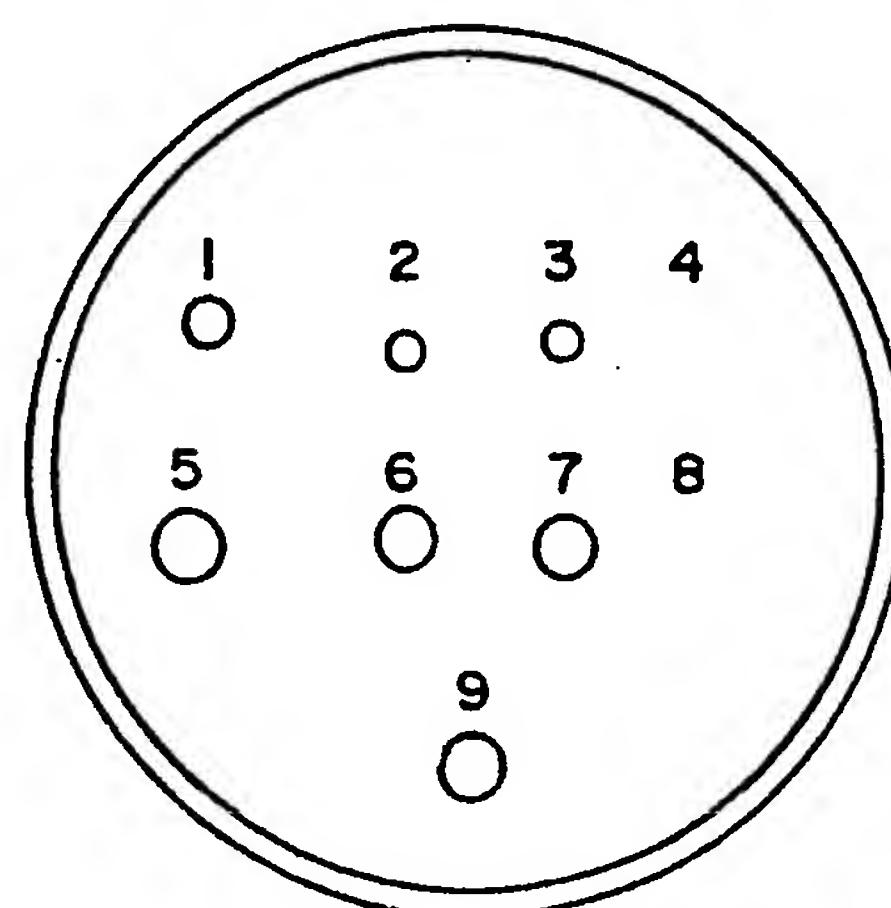
FIG. 14
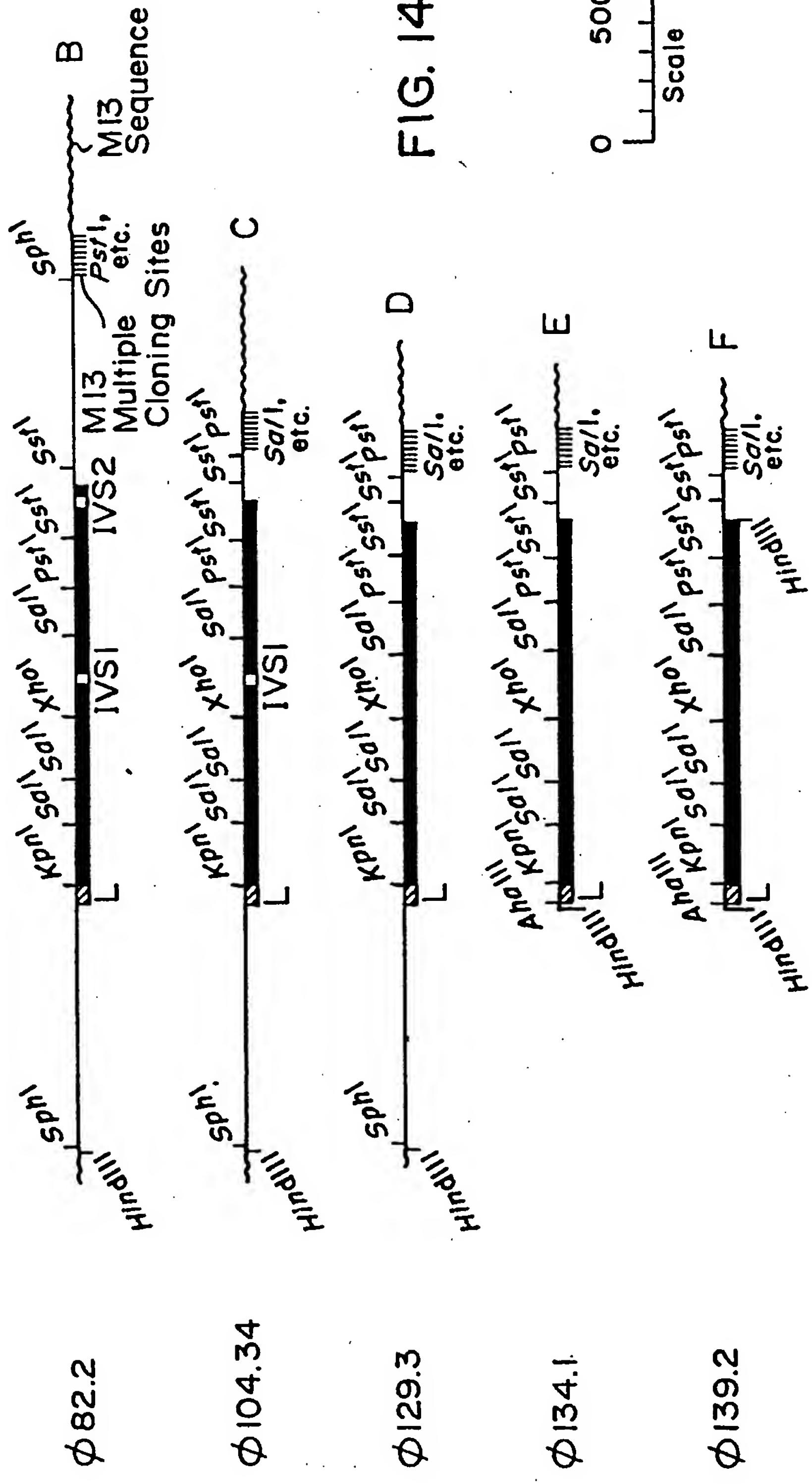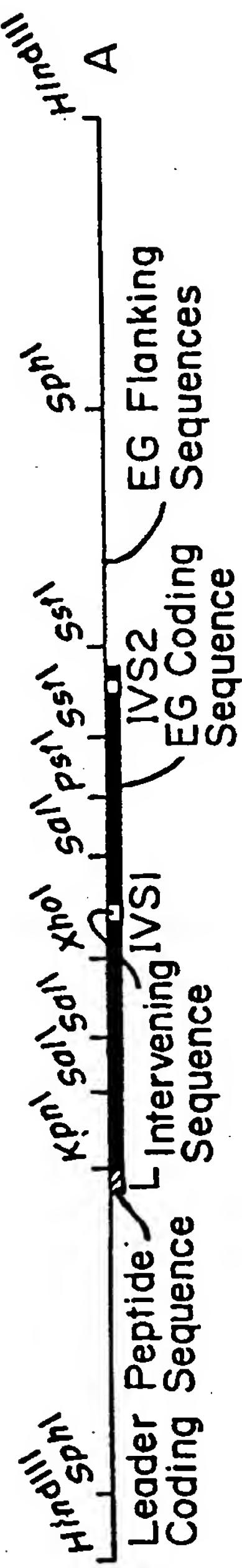
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int. Cl.4)
	DOCUMENTS CONSIDERED TO BE RELEVANT		EP 84110305.4
A,D	ABSTRACTS OF THE ANNUAL MEETING OF THE AMERICAN SOCIETY FOR MICROBIO-LOGY, 1981, 81st Annual Meeting, March 1-6, 1981 (Dallas, Texas) S.K. PICA TAGGIO et al. "The Cloning of Trichoderma, reesei Genomic DNA in Escherichia coli HB 101" page 116, column 2, Abstract H 19	1-20, 24-31	C 12 N 15/00 C 12 N 9/42 //C 12 R 1:885
A	EP - A1 - 0 011 767 (THE UNITED STATES OF AMERICA REPRESENTED BY THE SECRETARY U.S. DEPARTMENT OF COMMERCE) * Claims 1,2,6,7 *	1-3	
A,D	GENE, vol. 17, no. 1, January 1982 (Amsterdam) D.J. WHITTLE et al. "Molecular cloning of Cellulomonas fimi cellulase gene in Escherichia coli" pages 139-145 * Summary *	1	**TECHNICAL FIELDS SEARCHED (Int. Cl.4)** C 12 N

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
VIENNA	05-12-1984	WOLF

BEST AVAILABLE COPY

DECLARATION PURSUANT TO RULE 28, PARAGRAPH 4, OF THE EUROPEAN PATENT CONVENTION

The applicant has informed the European Patent Office that, until the publication of the mention of the grant of the European patent or until the date on which the application has been refused or withdrawn or is deemed to be withdrawn, the availability of the micro-organism(s) identified below, referred to in paragraph 3 of Rule 28 of the European Patent Convention, shall be effected only by the issue of a sample to an expert.

IDENTIFICATION OF THE MICRO-ORGANISMS

Accession numbers of the deposits:

A.T.C.C. 39766, 39767, 39768.

C.M.C.C. 1841, 1842, 1965, 2062, 2068.

N.R.R.L. 815573, 815574.

THIS PAGE BLANK (USPTO)